Crime Analytics: Visualization of Incident Reports

Tool: R with ggplot2

Question: Which incident types tend to correlate with each other on a day-by-day basis in SanFrancisco?

Figure 1 demonstrates changes of incident count by day that shows to days with more than 400 incidents and also 7 days with more than 350 incidents.

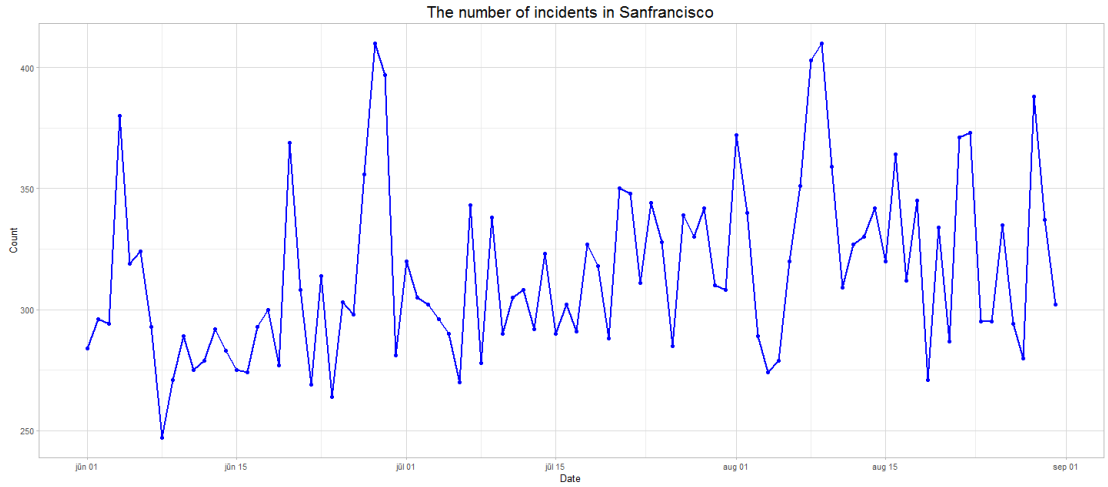

Figure 1

R code:

```
SanfData <- read.csv('…/sanfrancisco_incidents_summer_2014.csv')
SanfData[, "Date"] <- as.Date(SanfData[, "Date"], "%m/%d/%y")

Group= group_by(SanfData, Date)
Sanf_day_counts = summarise(Group, count = n())
ggplot(Sanf_day_counts, aes(Date, count, group = 1))+ geom_point(colour = "blue") +
  geom_line(colour = "blue", size = 1) +
  theme_light(base_size = 10)+ xlab("Date") + ylab("Count") +
  ggtitle("The number of incidents in Sanfrancisco") +
  theme(plot.title=element_text(size=16)) + scale_x_date(date_labels = "%b %d")
```

Figure 2 demonstrates average statistic of different incident types per day. Statistics shows that most often incident type is larceny/theft (average 100 incidents per day). Next most often incident types are ‘other offenses’, ‘assault’ and ‘non-criminal’.

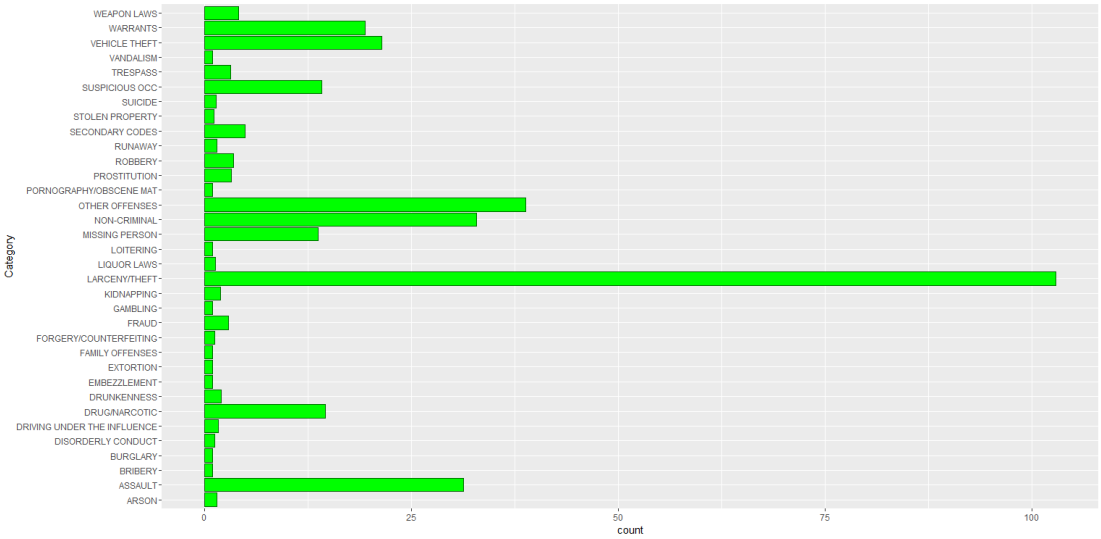

Figure 2

R code (cont.):

```
Group_categor = group_by(SanfData, Date, Category)
Sanf_day_counts2 = summarise(Group_categor, count = n())
Category_group = group_by(Sanf_day_counts2, Category)
Category_avg_counts = summarise(Category_group, count = mean(count))

ggplot(Category_avg_counts, aes(x = Category, y = count )) +
  geom_bar(colour="darkgreen", fill="green", stat = "identity") + coord_flip()
  theme_light(base_size = 12) + labs( y = "Category", x = "Count",) +
  ggtitle("The average number of incident by time of day") +
  theme(plot.title = element_text(size = 16))
```

Figure 3 demonstrates correlation between all incident types and Figure 4 between often incidents types. Statistics shows collation between count of all incidents and number larceny/theft incidents.

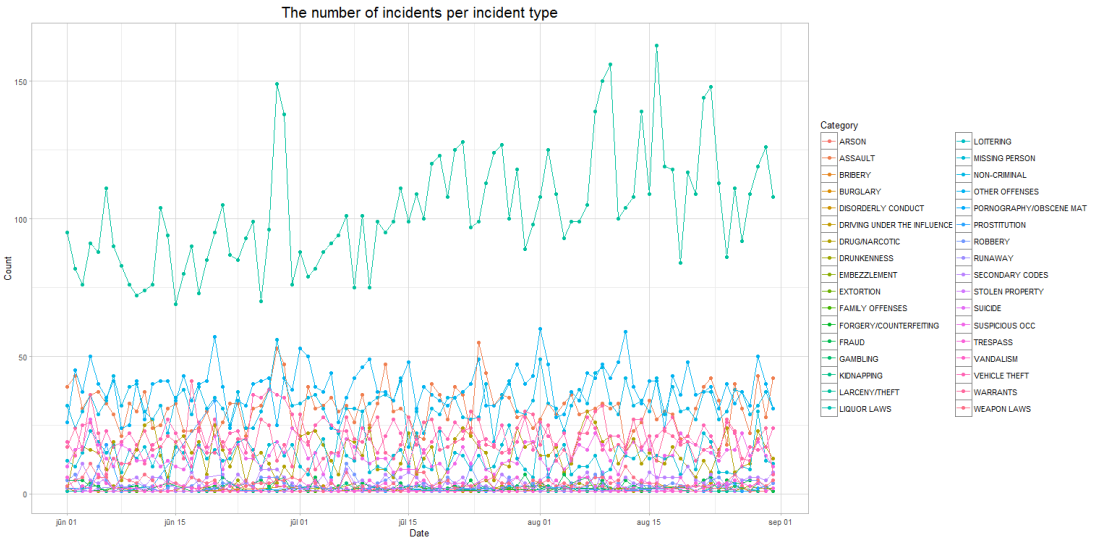

Figure 3

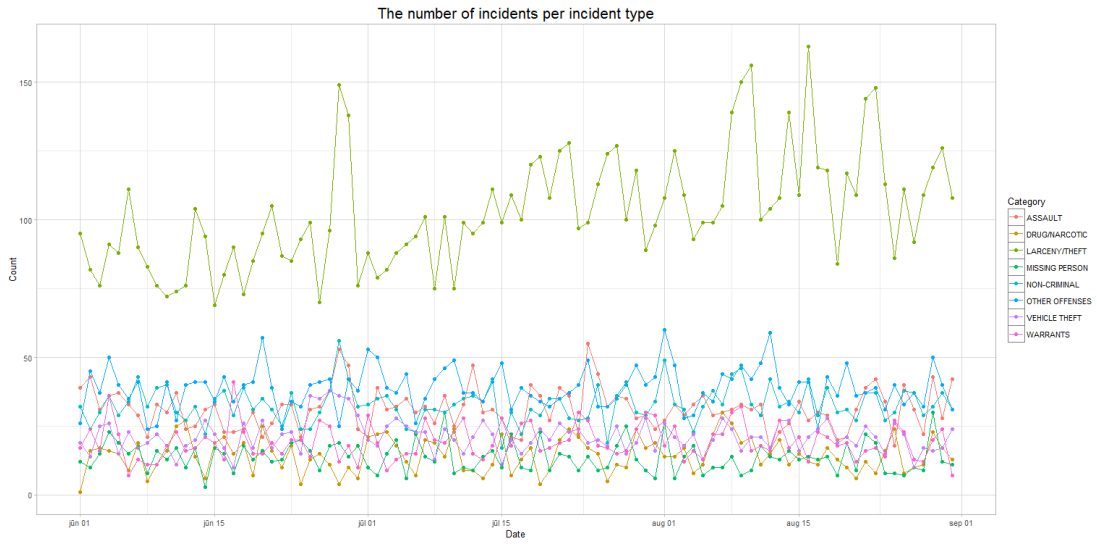

Figure 4

R code (cont.):

```
ggplot(Sanf_day_counts2, aes(Date, count, colour=Category))+ geom_point() +
  geom_line() +
  theme_light(base_size = 10)+ xlab("Date") + ylab("Count") +
  ggtitle("The number of incidents in Sanfrancisco") +
  theme(plot.title=element_text(size=16)) + scale_x_date(date_labels = "%b %d")
```

```
Sanf_day_counts3 = subset (Sanf_day_counts2, Category %in% c("LARCENY/THEFT",
"OTHER OFFENSES", "NON-CRIMINAL", "ASSAULT", "VEHICLE THEFT", "WARRANTS",
"DRUG/NARCOTIC", "MISSING PERSON"))
ggplot(Sanf_day_counts3, aes(Date, count, colour=Category))+ geom_point() +
  geom_line() +
  theme_light(base_size = 10)+ xlab("Date") + ylab("Count") +
  ggtitle("The number of incidents in Sanfrancisco") +
  theme(plot.title=element_text(size=16)) + scale_x_date(date_labels = "%b %d")
```

Figure 5 demonstrate correlation between categories and day of week. Statistic shows that more incidents in most often incident types are in Friday and Saturday.

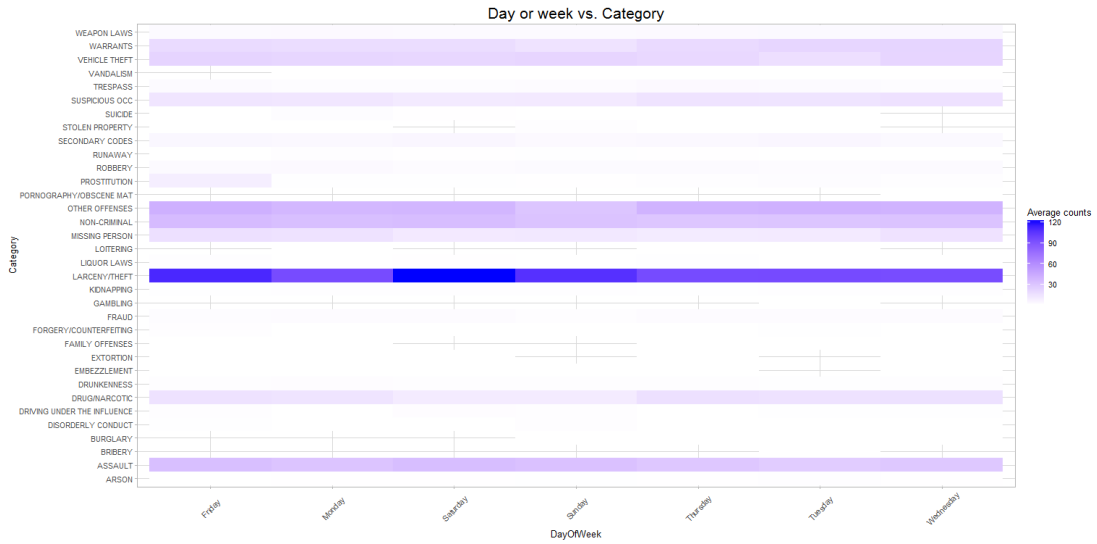

Figure 5

R code (cont.):

```
category_day = group_by(SanfData, Date, DayOfWeek, Category)
count_category_day = summarise(category_day, count = n())
category_dayofWeek = group_by(count_category_day, DayOfWeek, Category)
count_category_dayofWeek = summarise (category_dayofWeek,  CD_count =
mean(count))

ggplot(count_category_dayofWeek, aes(x = Category, y = DayOfWeek)) +
  geom_tile(aes(fill = CD_count)) +  coord_flip()+
  scale_fill_gradient(name="Average counts", low="white", high="blue") +
  theme(axis.title.y = element_blank()) + theme_light(base_size = 10) +
  theme(plot.title = element_text(size = 16)) +
  ggtitle("Day or week vs. Category")  +
  theme(axis.text.x = element_text(angle = 45,size = 8, vjust = 0.5))
```