**Descriptive Analysis**

**Student:***Kolonne Appuhamilage Charini Pubudika Jayasekera*

**Date Submitted:** *11/06/2022*

**Task Selected:** *Response Time*

Now that we have provided our initial project plan, and conducted a check on our data sources to assess potential sources of bias or error that could be introduced, we are now ready to begin the pathway to modeling. Before we can actually build a predictive model, we need to first get a feel for our data sets, the relationships between different variables, and existing patterns that could be relevant to our ultimate predictive model.

For this assignment, you will be constructing a series of descriptive analyses on both your primary data source (the EMS calls for service data) and any secondary data sources you are providing (whether they are weather data, Census data, etc.). Unlike our previous assignment, you will not need to do a separate analysis for each secondary dataset that would be introduced. In short, if you use the Community District dataset provided by the instructor, you can just run one set of descriptive analyses for the Community District dataset and one for the EMS calls for service. If you add weather data to your analysis, you will need to run a separate set of statistics for those data.

**WARNING**: This document will become very long (20 or so pages), primarily because you will be posting tables, charts, and then your own initial written assessments of what patterns you are seeing in the data. Please make sure you complete this assignment fully, as it will dramatically help your predictive modeling approach.

**Part 1 – EMS Calls for Service Data**

In this section, you will first analyze your calls for service data. This section is **required for everyone**, regardless of whether you are doing the demand modeling or the response time task. You will be providing descriptive tables and charts of the **individual calls for service data**, so no transformation is needed yet. We will move to that in Part 2a for those doing demand modeling, while those doing the response time will be working on Part 2b.

For this section, you will need to complete the following tables, charts, and assessments. There is a space on each page for these items, and you will need to provide not only the formatted items, but also a short (2-3 sentences) description of what patterns you are seeing. Please conduct the requested analyses, insert the formatted items, and complete the descriptions. Make sure the formatting stays the same, so that the presentation is consistent across pages.

All the relevant code to create tables and charts you can find in the training\_data\_load.r file on the Blackboard page.

* - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -- - - - - - - - - - - - - - - - - - - - - - - - - -

***# of calls by year***

Chart, bar chart

Description automatically generated

# bar chart calls by the year

counts\_year <- table(a$incident\_year)

barplot(counts\_total, xlab="Year", ylab="# of Calls", col=c("red"), main = "EMS Calls by Year")

What patterns are you seeing?

*Based on the above bar graph we can see the increment of the calls and incidents over the years, it could be a true or a fake incident but overall, the calls have been increased. Since 2008 there had been an increment, but it was almost equal in the years of 2015 and 2016, the amounts of calls received and recorded. But overall it was a gradual increment not a sudden increment over the years. Therefore patients and accidents have got increased over the years.*

***# of calls by day of week***

Chart, bar chart

Description automatically generated

# bar chart calls by the day of the week

counts\_days <- table(a$dow)

barplot(counts\_days, xlab="DOW", ylab="# of Calls", col=c("red"), main = "EMS Calls by days of the week")

What patterns are you seeing?

*We can see that during the weekends the calls and incidents are less. Specially on Sundays. A reason for this could be that no work during the weekends for most of the people therefore less travel could cause less traffic accidents. Which can reduce many calls for the weekend as people tend to stay at their house. But during the weekdays it had increased equally in all the 5 days. Therefore we can identify that weekdays have more incidents than weekends.*

***# of calls by month (i.e. Jan-Dec)***

Chart, bar chart

Description automatically generated

# bar chart calls by the month

counts\_calls\_monthly <- table(a$month)

barplot(counts\_calls\_monthly, xlab="Month", ylab="# of Calls", col=c("red"), main = "EMS Calls by Month")

A screenshot of a computer

Description automatically generated with medium confidence

data.frame(table(a$month))

Text

Description automatically generated

What patterns are you seeing?:

*A higher rate of incidents can be found in months like May, June, July due to summer time. Summer time is the periods from May – August where people travel a lot and go for outdoor activities and can have many accidents. But also in the month of December people go out to see snow and Christmas with parties can increase the accidents. People get into unexpected fights, crime scenes can increase during the gatherings, parties and can cause traffic accidents while travelling which have shown a higher incident rate in December month. And even in October a higher rate could be seen as it is the beginning of Fall season. People tend to go trips during the Fall period beginning as well to see the fall colors. Hence a higher rate of incidents can be seen in summer months, October and December.*

***# of calls by hour of day (i.e. 12am-12am)***

Chart, bar chart, histogram

Description automatically generated

#create hourly variable

a$hourly <- format(as.POSIXct(a$incident\_dt), format = "%H")

# bar chart calls by the month

counts\_calls\_hourly <- table(a$hourly)

barplot(counts\_calls\_hourly, xlab="Hour", ylab="# of Calls", col=c("red"), main = "EMS Calls Hourly")

What patterns are you seeing?:

*A higher rate of incidents can be seen during the 11 am to 4 pm of the day. This could be due to many go to office work during the day time and school hours are during that time as well. These can have road accidents during this time of the day. Hence it has got increased.*

***# of calls by month and year (i.e. Jan 2008 – Dec 2016)***

Chart

Description automatically generated

What patterns are you seeing?

*In this chart we can see that the incidents reported got increased in the months and years of Aug 2009, Dec 2010, Jul 2012, Jun 2012, Mar 2013, May 2014, Nov 2015 and Oct 2016. As mentioned previously these months are either summer months or Christmas, Fall beginning months which have a higher rate of incidents.*

***# of calls by date (i.e. 1/1/2008, 1/2/2008, etc.)*** ***Chart, shape

Description automatically generated***

What patterns are you seeing?

*In this chart it’s shown the highest incident dates in the graph and it has a similar point of view I mentioned about summer dates, Fall time and Christmas time periods that can have many accidents.*

***# of calls by Borough***Chart, bar chart

Description automatically generated

# bar chart calls by the month

counts\_calls\_borough <- table(a$borough)

barplot(counts\_calls\_borough, xlab="Borough", ylab="# of Calls", col=c("red"), main = "EMS Calls by Borough")

Text

Description automatically generated

data.frame(table(a$borough))

What patterns are you seeing?

*Usually highest population can be seen in 4 boroughs apart from Staten Island. Hence Comparatively Staten Island has the lowest incident rates. The rest of the areas have a similar incident rate over the years compared to those total population.*

***# of calls by Community District***

# table calls by the community district

counts\_calls\_community\_dis <- table(a$communitydistrict)

data.frame(table(a$communitydistrict))

Table

Description automatically generated A close-up of a chart

Description automatically generated with low confidence

What patterns are you seeing?

*Community districts have different incident rates based on many factors as the number of population, crime rate in those areas.*

***# of calls by Initial Call Type***

# table calls by the initial call type

table(a$initial\_call\_type) %>% as.data.frame() %>% arrange(desc(Freq))

Table

Description automatically generated with low confidenceA black and white document

Description automatically generated with low confidenceTable

Description automatically generated with medium confidence

What patterns are you seeing?

*We can see that there is a high call rate for Sick, Non-critical Injury, Difficult breather and the lowest calls called for shooters, civil disturbance, aircraft incident crashes etc. This can be seen as getting sick can be a general problem including calling for any small injury for ambulances. But we rarely get aircraft crashing, report of explosives which can be seen with less calls.*

***# of calls by Final Call Type***

# table calls by the final call type

table(a$final\_call\_type) %>% as.data.frame() %>% arrange(desc(Freq))

A picture containing table

Description automatically generatedA picture containing text, newspaper, receipt

Description automatically generated

What patterns are you seeing?

*Final call type also can be seen with the same way of the initial call type pattern but with a change of numbers. Categorizing the initially identified call type to a more discreet final correct call category.*

**# of calls, by Initial Severity Level and Final Severity Level**

req <- substitute(require(x, character.only = TRUE))

libs<-c("sjPlot")

#sapply(libs, function(x) eval(req) || {install.packages(x); eval(req)})

sjPlot::tab\_xtab(var.row = a$initial\_call\_type, var.col = a$final\_call\_type, title = "Table Title", show.row.prc = TRUE)

Scatter chart

Description automatically generated with low confidence A picture containing chart

Description automatically generated

What patterns are you seeing?

*cross-tabulations are simply data tables that present the results of the entire group of respondents, as well as results from subgroups of survey respondents. With them, we can examine relationships within the data that might not be readily apparent when only looking at total dataset. Based on this dataset we can see how final dataset and initial call dataset is matching with each other.*

**Part 2 – Task Analysis**

Now that we have completed some initial descriptive analyses of the individual calls for service data, it is time to move onto our specific sections related to the task that we have chosen. In this section, you will complete either Part 2a (for those doing the demand modeling) or Part 2b (for those doing the response time analysis).

For the part you are not doing, please delete that part from your final submission.

Please complete all sections, and ensure that the analyses provided are formatted, well-described, and clear.

**Part 2b – Response Time Analysis**

For those who are doing the response time analysis, please complete the following descriptive statistic tables and/or charts. After each group of charts or tables, I want you to include the following section:

Here are the charts to complete:

1. Distribution of calls by length of incident\_response\_seconds\_qy

#bar chart of incidents by Response Time

counts\_total <- table(a$incident\_response\_seconds\_qy)

barplot(counts\_total, xlab="Seconds", ylab="# of Incidents", col=c("red"), main = "EMS Incidents by Response Seconds")

#line chart of incidents by Response Time

plot(counts\_total, type="o", xlab="Seconds", ylab="# of Incidents", col=c("red"), main = "EMS Incidents by Response Seconds")

data.frame(table(a$incident\_response\_seconds\_qy))

table(a$incident\_response\_seconds\_qy) %>% as.data.frame() %>% arrange(desc(Freq))

**Graphical user interface, text, application

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

What patterns are you seeing?:

*I can see there are many incidents around first 15 mins of a call length in this scenario with a maximum of*

1. Distribution of calls by length of dispatch\_response\_seconds\_qy

#bar chart of dispatch by Response Time

counts\_total <- table(a$dispatch\_response\_seconds\_qy, a$incident\_dt)

barplot(counts\_total, xlab="Seconds", ylab="# of Incidents", col=c("red"), main = "EMS Dispatch by Response Seconds")

#line chart of incidents by Response Time

plot(counts\_total, type="o", xlab="Seconds", ylab="# of Incidents", col=c("red"), main = "EMS Dispatch by Response Seconds")

data.frame(table(a$dispatch\_response\_seconds\_qy,a$incident\_dt))

table(a$dispatch\_response\_seconds\_qy) %>% as.data.frame() %>% arrange((Freq))

**Graphical user interface, text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated with medium confidence**

*In this we can see that there are few outliers which can have an impact in the dataset, but also we can see that the final dispatch has been done quickly as possible.*

1. Distribution of calls by length of incident\_travel\_tm\_seconds\_qy

counts\_calls\_travel\_tm <- table(a$incident\_travel\_tm\_seconds\_qy)

barplot(counts\_calls\_travel\_tm, xlab="", ylab="Travel\_TM", col=c("red"), main = "EMS Calls travel tm seconds")

data.frame(table(a$incident\_travel\_tm\_seconds\_qy))

table(a$incident\_travel\_tm\_seconds\_qy) %>% as.data.frame() %>% arrange(desc(Freq))

**Graphical user interface, text, application

Description automatically generated**

**Text

Description automatically generated**

*Based on these data we can see that the first assigned time and the first on the scene time has the biggest time gap as 7 hours and the rest have a time gap less than that.*

1. Distribution of calls by length of incident\_response\_seconds\_qy, for each Borough (i.e. five separate distributions)

counts\_calls\_borough <- table(a$incident\_response\_seconds\_qy, a$borough)

barplot(counts\_calls\_borough, xlab="Borough", ylab="Responded Seconds", col=c("red"), main = "EMS Calls Responded Seconds by Borough")

data.frame(table(a$incident\_response\_seconds\_qy, a$borough))

table(a$incident\_response\_seconds\_qy,a$borough) %>% as.data.frame() %>% arrange(desc(Freq))

**Chart, bar chart

Description automatically generatedText

Description automatically generated with medium confidence**

*We can see that Brooklyn has the highest incidents and the lowest response time as well compared to other boroughs. Brooklyn takes the maximum time to go to the incident place as 2 hours and respond to emergency services compaed to other boroughs where as Staten Island EPS takes the lowest time to respond.*

1. Distribution of calls by length of incident\_response\_seconds\_qy, for day of the week (i.e. seven separate distributions)

**Chart, bar chart

Description automatically generatedA picture containing text

Description automatically generated**

*Response time to go to an incident place is taken in weekends than weekdays. Sunday takes the lowest response time due to less traffic on that day to travel to the incident place. But the rest of the days take a similar time to respond to go to an incident place.*

1. Distribution of calls by length of incident\_response\_seconds\_qy, for each month (i.e. 12 separate distributions)

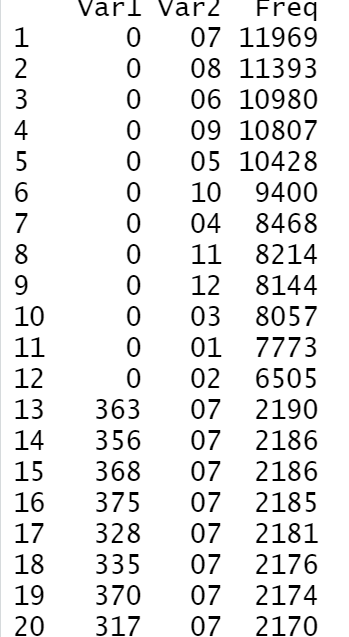
counts\_calls\_inci\_resp\_month <- table(a$incident\_response\_seconds\_qy,a$month)

barplot(counts\_calls\_inci\_resp\_month, xlab="Month", ylab="Incidents\_Res\_Seconds", col=c("red"), main = "EMS Calls Inci Resp Time Month")

table(a$incident\_response\_seconds\_qy,a$month) %>% as.data.frame() %>% arrange(desc(Freq))

Chart, bar chart, histogram

Description automatically generated

**

*It can see that low response time is taken by less function months such as February, April, September, November. This is due to less number of incidents during those months. But in the months where are there are more incidents such as summer months, Christmas and new year months have taken more time to respond due to high number of incidents in those months. This is the time elapsed in seconds between the incident\_datetime and the first\_on\_scene\_datetime.*

1. Distribution of calls by length of incident\_travel\_tm\_seconds\_qy, for each month (i.e. 12 separate distributions)

counts\_calls\_inci\_travel\_tm\_month <- table(a$incident\_travel\_tm\_seconds\_qy,a$month)

barplot(counts\_calls\_inci\_travel\_tm\_month, xlab="Month", ylab="Incidents\_Travel\_tm", col=c("red"), main = "EMS Calls Inci Travel Month")

table(a$incident\_travel\_tm\_seconds\_qy,a$month) %>% as.data.frame() %>% arrange(desc(Freq))

Chart, bar chart, histogram

Description automatically generatedText

Description automatically generated

*The time elapsed in seconds between the first\_assignment\_datetime and the first\_on\_scene\_datetime also have a similar pattern as in incidents response call seconds pattern as above. During less busier months they have taken less time to respond where as more busier months have taken longer to respond.*

**Part 3 – Additional Data Analysis**

Now that you have completed your analysis of the overall data and your specific task requirements, it is time to focus on the additional data you are planning on using for the modeling. This could be weather data, this could be community district data, or other data sources.

In the following section, you are to conduct **1 additional table or chart (your choice) for each variable of interest** you plan on using from these alternative datasets. So, if you plan on using weather data (like precipitation), you will need to create a chart that shows precipitation by month of year, and/or precipitation by year. If you are using the Community District data, you need to do a chart or table for each variable you use.

You should have several charts/tables in this section, as everyone should be using additional data in their modeling approach.

Finally, after including all the relevant charts and tables, I want you to spend 1-2 paragraphs describing the additional patterns that you see in these data, and how they might be related to your dependent variable of interest (# of calls or response time).

**Additional Dataset 1 – 911\_Open\_Data\_Local\_Law\_119**

library(expss)

setwd("C:/Users/sachi/OneDrive - University at Albany - SUNY/SUNY/Courses/Semester 04 - Fall 22/CINF 624/Assignment/My Own Data")

getwd()

b <-read.csv("911\_Open\_Data\_Local\_Law\_119.csv")

head(b)

library(data.table)

library(ggplot2)

library(reshape)

library(lubridate)

library(dplyr)

**#Calls by the Borough**

data.frame(table(b$X..of.Incidents, b$Borough))

counts\_calls\_borough <- table(b$X..of.Incidents, b$Borough)

barplot(counts\_calls\_borough, xlab="Borough", ylab="# of Calls", col=c("red"), main = "Calls by Borough")

**Background pattern

Description automatically generated**

**#Calls by the Month**

# bar chart calls by the Month

counts\_calls\_month <- table(b$X..of.Incidents, b$Month.Name)

barplot(counts\_calls\_month, xlab="Month", ylab="# of Calls", col=c("red"), main = "Calls by Month")

data.frame(table(b$X..of.Incidents, b$Month.Name))

**Chart

Description automatically generated**

**#calls by weekdays**

library("anytime")

# bar chart calls by weekdays

b$dow <- weekdays(anydate(b$X..of.Incidents))

counts\_days <- table(b$X..of.Incidents,b$dow)

barplot(counts\_days, xlab="DOW", ylab="# of Calls", col=c("red"), main = "Calls by days of the week")

Chart, bar chart

Description automatically generated

**#calls by Agency**

data.frame(table(b$X..of.Incidents, b$Agency))

counts\_calls\_agency <- table(b$X..of.Incidents, b$Agency)

barplot(counts\_calls\_agency, xlab="Agency", ylab="# of Calls", col=c("red"), main = "Calls by Agency")

Chart

Description automatically generated with low confidence

**#calls by description**

data.frame(table(b$X..of.Incidents, b$Description))

counts\_calls\_desc <- table(b$X..of.Incidents, b$Description)

barplot(counts\_calls\_desc, xlab="Description", ylab="# of Calls", col=c("red"), main = "Calls by Description")

Chart, diagram, bar chart

Description automatically generated

**Response Time by DOW**

data.frame(table(b$Response.Times, b$dow))

counts\_calls\_Resp <- table(b$Response.Times, b$dow)

barplot(counts\_calls\_Resp, xlab="DOW", ylab="Response Times", col=c("red"), main = "Response Time by DOW")

**Chart, bar chart

Description automatically generated**

**Response Time by Borough**

data.frame(table(b$Response.Times, b$Borough))

counts\_calls\_Resp <- table(b$Response.Times, b$Borough)

barplot(counts\_calls\_Resp, xlab="Borough", ylab="Response Times", col=c("red"), main = "Response Time by Borough")

**Chart, bar chart

Description automatically generated**

**Response Time by Agency**

data.frame(table(b$Response.Times, b$Agency))

counts\_calls\_Resp <- table(b$Response.Times, b$Agency)

barplot(counts\_calls\_Resp, xlab="Agency", ylab="Response Times", col=c("red"), main = "Response Time by Agency")

**A picture containing chart

Description automatically generated**

**Additional Dataset 2 – NYC\_Health\_\_\_Hospitals\_patient\_care\_locations\_-\_2011**

library(expss)

getwd()

setwd("C:/Users/sachi/OneDrive - University at Albany - SUNY/SUNY/Courses/Semester 04 - Fall 22/CINF 624/Assignment/My Own Data")

c <-read.csv("NYC\_Health\_\_\_Hospitals\_patient\_care\_locations\_-\_2011.csv")

head(b)

library(data.table)

library(ggplot2)

library(reshape)

library(lubridate)

library(dplyr)

Facilities\_Type <- table(c$Facility.Type)

plot(Facilities\_Type, xlab="Facility Name", ylab="# of Incidents Per Facility", col=c("red"), main = "Incidents by Facility")

Chart

Description automatically generated

Borough <- table(c$Borough)

plot(Borough, xlab="Borough", ylab="# of Incidents", col=c("red"), main = "Incidets by Borough")

Chart

Description automatically generated

What patterns are you seeing in this additional data you are including in the model?

*Additional Dataset 1 –*

*In Additional Dataset 1 is based on 911 open data calls and I have done the graphs based on Borough, Days, Agency Name, Month and Response Time. When we take the days graph, we can see Fridays have the highest calls to the agencies. This can be probably most of the people go on trips, home after rush could cause many accidents hence it can have many calls to 911. Also, when people go trips on weekends can cause higher calls amount.*

*When we look at the agency graph, we can interpret that FDNY answers the highest rates of calls and handle many incidents. Which means most of the 911 calls are gone to FDNY agency than the other two agencies, EMS and Other. At the same time, we can see that it has a positive correlation with response time as well with the highest response rate in Fridays and by FDNY agency.*

*Additional Dataset 2 –*

*In this dataset I have done graphs based on Facility Name and Borough. Based on this dataset we can see which boroughs have the most hospital locations and what facilities get the highest patient records from 2011 to 2022. Based on this we can see Child health center has more incidents and Brooklyn has the highest cases for these facilities. Whereas Diagnostic and Treatment center and Nursing homes have the lowest cases and at the Staten Island with the lowest cases. Therefore, this dataset helps to identify the NYC patient care locations for hospitals.*

How might these additional factors be related to your modeling task (either # of calls or response time)?

*I have selected the response time as my main task to model. Hence the number of calls and time taken to response to those calls are a major factor that can impact the number of incidents.*

*The additional factors such as the facility names and the facility locations can help to know which healthcare facilities closer to each borough’s exact incident location in NY. Therefore I have taken that dataset.*

*Also We can see which agency responses on the calls of 911, quickly whether it’s FDNY, EMS and etc. Therefore we can predict that these agencies quickly responses for future as well. Also the days that 911 gets more calls such as Fridays which will help them to keep more call centers active for these peak days and for the boroughs. Likewise we can predict early the higher incident areas early from these additional datasets for us to work on these future and keep more active callers and emergency care on those specific categories such as highest/peak days of incidents, times, boroughs, facilities and etc. Hence these additional datasets help to predict details for the main dataset.*