Business Statistics Mid-Term Assessment IB94X0 2022-2023 #1

Sirapob Lurojruang - 2215107

This is to certify that the work I am submitting is my own. All external references and sources are clearly acknowledged and identified within the contents. I am aware of the University of Warwick regulation concerning plagiarism and collusion.

No substantial part(s) of the work submitted here has also been submitted by me in other assessments for accredited courses of study, and I acknowledge that if this has been done an appropriate reduction in the mark I might otherwise have received will be made

---

```
library(tidyverse)
library(ggplot2)
library(lubridate)
library(grid)
library(gridExtra)
library(knitr)
library(emmeans)
options(width=100)
```

Section 1

Importing data and dictionary

Variables	Description
ProperCase	Borough name with a proper case

Variables	Description
FirstPumpArriving_AttendanceTime	The attendance time (in seconds) for the first fire engine to arrive after it has been mobilised from a fire station (or other location if it was mobile by Brigade Control at the time of the call). When fire crews arrive they record their attendance using an on-board computer (a Mobile Data Terminal). There will be occasions when the first crew to arrive fail to record this correctly (either as human error or a delay/failure in the communications). When this happens the time recorded may in fact be the second or third.
Notional.Cost..Â..	An estimate of the cost of the incident response
cost_responding	Duplicate of Notional.Cost..Â.. created for better clarity

```
#Import data
fire_data_raw <- read.csv("London_Fire_data.csv")

# There are some outlier in which skew the mean of the data as indicated by significantly large
max and difference in mean and median of cost and response time.
summary(fire_data_raw)
```

```
##  IncidentNumber      DateOfCall           CalYear      TimeOfCall          HourOfCall
##  Length:322375      Length:322375      Min.   :2019   Length:322375      Min.   : 0.00
##  Class :character   Class :character   1st Qu.:2019   Class :character   1st Qu.: 9.00
##  Mode  :character   Mode  :character   Median :2020   Mode  :character   Median :14.00
##                                        Mean   :2020                      Mean   :13.42
##                                        3rd Qu.:2021                      3rd Qu.:19.00
##                                        Max.   :2022                      Max.   :23.00
##
##  IncidentGroup      StopCodeDescription SpecialServiceType PropertyCategory   PropertyType
##  Length:322375      Length:322375       Length:322375      Length:322375      Length:322375
##  Class :character   Class :character    Class :character   Class :character   Class :characte
r
##  Mode  :character   Mode  :character    Mode  :character   Mode  :character   Mode  :characte
r
##
##
##
##
##  AddressQualifier   Postcode_full      Postcode_district       UPRN                 USRN
##  Length:322375      Length:322375      Length:322375      Min.   :0.000e+00   Min.   : 420074
0
##  Class :character   Class :character   Class :character   1st Qu.:0.000e+00   1st Qu.:2040098
9
##  Mode  :character   Mode  :character   Mode  :character   Median :0.000e+00   Median :2120112
1
##                                                           Mean   :2.072e+10   Mean   :2040083
7
##                                                           3rd Qu.:1.001e+10   3rd Qu.:2210081
3
##                                                           Max.   :2.000e+11   Max.   :9999042
2
##
##  IncGeo_BoroughCode IncGeo_BoroughName  ProperCase         IncGeo_WardCode    IncGeo_WardName
##  Length:322375      Length:322375      Length:322375      Length:322375      Length:322375
##  Class :character   Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  IncGeo_WardNameNew   Easting_m        Northing_m     Easting_rounded  Northing_rounded
##  Length:322375      Min.   :503582   Min.   :155998   Min.   :503550   Min.   :155950
##  Class :character   1st Qu.:524924   1st Qu.:175804   1st Qu.:525150   1st Qu.:176050
##  Mode  :character   Median :530858   Median :180978   Median :530950   Median :181050
##                     Mean   :530634   Mean   :180340   Mean   :530667   Mean   :180487
##                     3rd Qu.:537035   3rd Qu.:185076   3rd Qu.:536350   3rd Qu.:185250
##                     Max.   :560461   Max.   :200885   Max.   :611150   Max.   :302450
##                     NA's   :175667   NA's   :175667
##     Latitude        Longitude           FRS            IncidentStationGround
##  Min.   : 0.00    Min.   :-0.51    Length:322375      Length:322375
##  1st Qu.:51.47    1st Qu.:-0.20    Class :character   Class :character
##  Median :51.51    Median :-0.12    Mode  :character   Mode  :character
```

```
##  Mean   :51.36    Mean   :-0.12
##  3rd Qu.:51.55    3rd Qu.:-0.03
##  Max.   :51.69    Max.   : 0.31
##  NA's   :175667   NA's   :175667
##  FirstPumpArriving_AttendanceTime FirstPumpArriving_DeployedFromStation
##  Min.   :   1.0                   Length:322375
##  1st Qu.: 227.0                   Class :character
##  Median : 290.5                   Mode  :character
##  Mean   : 308.1
##  3rd Qu.: 367.0
##  Max.   :1199.0
##  NA's   :19019
##  SecondPumpArriving_AttendanceTime SecondPumpArriving_DeployedFromStation
##  Min.   :   1.0                    Length:322375
##  1st Qu.: 293.0                    Class :character
##  Median : 363.0                    Mode  :character
##  Mean   : 385.6
##  3rd Qu.: 450.0
##  Max.   :1200.0
##  NA's   :199385
##  NumStationsWithPumpsAttending NumPumpsAttending   PumpCount       PumpHoursRoundUp
##  Min.   : 1.0                  Min.   : 1.000    Min.   :  1.000   Min.   :   1.00
##  1st Qu.: 1.0                  1st Qu.: 1.000    1st Qu.:  1.000   1st Qu.:   1.00
##  Median : 1.0                  Median : 1.000    Median :  1.000   Median :   1.00
##  Mean   : 1.4                  Mean   : 1.571    Mean   :  1.619   Mean   :   1.37
##  3rd Qu.: 2.0                  3rd Qu.: 2.000    3rd Qu.:  2.000   3rd Qu.:   1.00
##  Max.   :14.0                  Max.   :14.000    Max.   :250.000   Max.   :1203.00
##  NA's   :3823                  NA's   :3823      NA's   :2008      NA's   :2111
##  Notional.Cost..Â..    NumCalls
##  Min.   :   333.0   Min.   :  1.000
##  1st Qu.:   339.0   1st Qu.:  1.000
##  Median :   346.0   Median :  1.000
##  Mean   :   471.9   Mean   :  1.306
##  3rd Qu.:   352.0   3rd Qu.:  1.000
##  Max.   :407817.0   Max.   :175.000
##  NA's   :2111       NA's   :4
```

The cost of response time

Preparing Data

```
# Check distribution of cost data with histogram. The result shows that there are multiple rare
costly incidents which skew the mean of the data.
grid.arrange(
  ggplot(fire_data_raw, aes(Notional.Cost..Â..)) +
    geom_histogram(binwidth = 1000) +
    scale_y_log10() +
    labs(title = "Distribution of cost (Y log 10)", x = "Notional Cost", y = "Frequency (log 1
0)"),
  ggplot(fire_data_raw, aes(Notional.Cost..Â..)) +
    geom_histogram(binwidth = 1000) +
    facet_grid(IncidentGroup~.) +
    xlim(0,100000) +
    scale_y_log10() +
    labs(title = "Distribution of cost by Incident (Y log 10)", x = "Notional Cost (xlim 100k)",
y = "Frequency (log 10)"),
ncol = 2)
```

```
## Warning: Removed 2111 rows containing non-finite values (stat_bin).
```

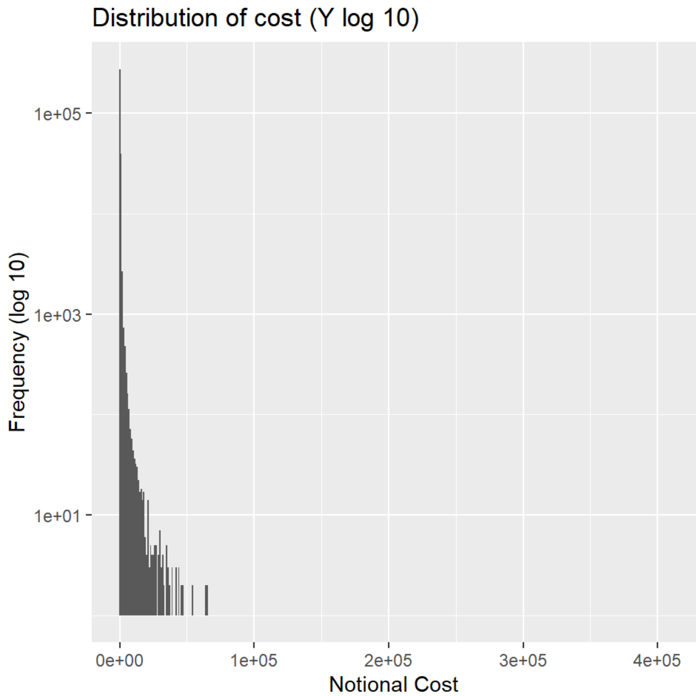
```
## Warning: Transformation introduced infinite values in continuous y-axis
```

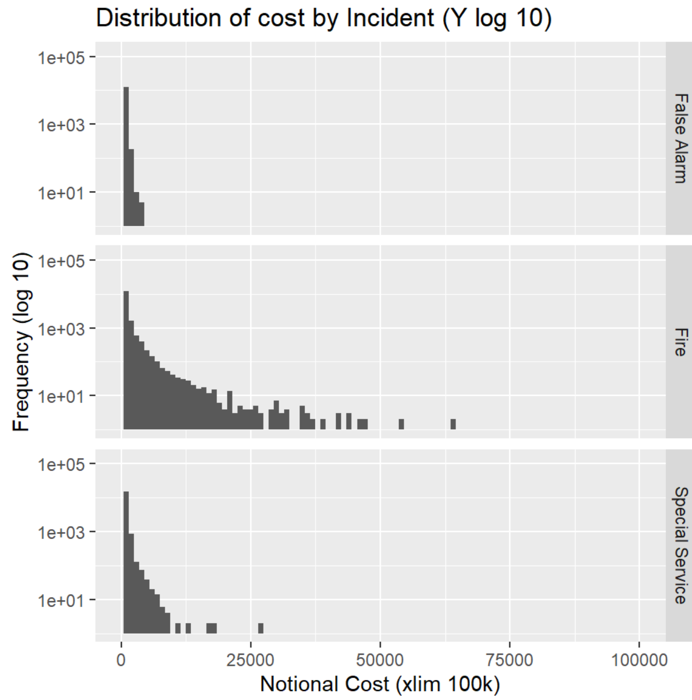
```
## Warning: Removed 323 rows containing missing values (geom_bar).
```

```
## Warning: Removed 2131 rows containing non-finite values (stat_bin).
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
```

```
## Warning: Removed 212 rows containing missing values (geom_bar).
```






```
# While IQR method of Outlier detection was considered, the right-hand side split by Incident ch
art above indicated that majority of rare costly incidents are concentrated in Fire incident whi
ch will be removed should the IQR method applied.

# Looking at the distribution, the continuity of the distribution seems to end around 100k mark.
Hence, the cut-off point of 100k is apply to the dataset.
```

```
# Apply upper bound of 100k to the dataset to exclude outlier
lower_bound_cost <- 0
upper_bound_cost <- 100000

fire_data_cost <- fire_data_raw %>%
  mutate(cost_responding = Notional.Cost..Â..) %>% # Mutate Notional cost column to cost_respond
ing for clarity and cleaner codes.
  filter(cost_responding <= upper_bound_cost, cost_responding >= lower_bound_cost)

# Mean of cost decreased by 10.90 after clear outliers
mean(fire_data_raw$Notional.Cost..Â.., na.rm = TRUE) - mean(fire_data_cost$cost_responding, na.r
m = TRUE)
```

```
## [1] 10.90732
```

Calculating cost of each fires

```
# Filter incidents to excluded Special Services
(respond_cost_type <- fire_data_cost %>%
  filter(IncidentGroup != "Special Service") %>%
  group_by(IncidentGroup) %>%
  summarise(total_cost = sum(cost_responding, na.rm = TRUE),
            avg_cost = mean(cost_responding, na.rm = TRUE)))
```

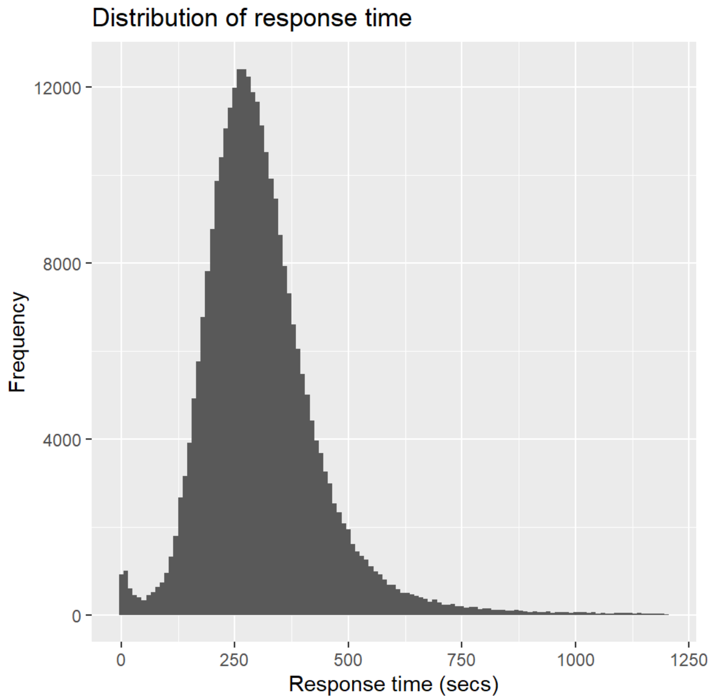
```
## # A tibble: 2 x 3
##   IncidentGroup total_cost avg_cost
##   <chr>              <int>    <dbl>
## 1 False Alarm     61249812     378.
## 2 Fire            39676816     772.
```

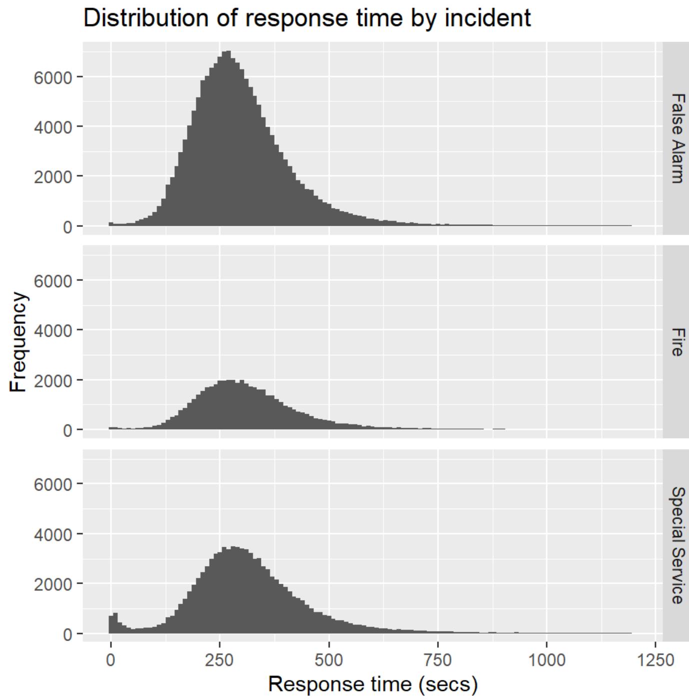
The distribution of response times

Preparing Data

```
# Check distribution of cost data with histogram. While the distribution of response time seems
to be normally distributed, the chart below illustrate that the data is right-skewed.
grid.arrange(
  ggplot(fire_data_raw, aes(FirstPumpArriving_AttendanceTime)) +
    geom_histogram(binwidth = 10) +
    labs(title = "Distribution of response time", x = "Response time (secs)", y = "Frequency"),
  ggplot(fire_data_raw, aes(FirstPumpArriving_AttendanceTime)) +
    geom_histogram(binwidth = 10) +
    facet_grid(IncidentGroup~.) +
    labs(title = "Distribution of response time by incident", x = "Response time (secs)", y = "F
requency"),
ncol = 2)
```

```
## Warning: Removed 19019 rows containing non-finite values (stat_bin).
## Removed 19019 rows containing non-finite values (stat_bin).
```






```
# Since there is a long-tail of data points, IQR method is an appropiate method to prepare the d
ata for further analysis.
```

```
# Use Inter Quartile Range to determine outlier range
time_q1 <- quantile(fire_data_raw$FirstPumpArriving_AttendanceTime, probs = 0.25, na.rm = TRUE)
time_q3 <- quantile(fire_data_raw$FirstPumpArriving_AttendanceTime, probs = 0.75, na.rm = TRUE)
IQR_time <- time_q3 - time_q1
upper_bound_time <- time_q3 + (1.5*IQR_time)
lower_bound_time <- time_q1 - (1.5*IQR_time)

# Apply upper and lower bound to the dataset to exclude outlier
fire_data_time <- fire_data_raw %>%
  filter(FirstPumpArriving_AttendanceTime <= upper_bound_time, FirstPumpArriving_AttendanceTime
>= lower_bound_time)

# Mean of response time decreased by 12.50 after clear outliers
mean(fire_data_raw$FirstPumpArriving_AttendanceTime, na.rm = TRUE) - mean(fire_data_time$FirstPu
mpArriving_AttendanceTime, na.rm = TRUE)
```

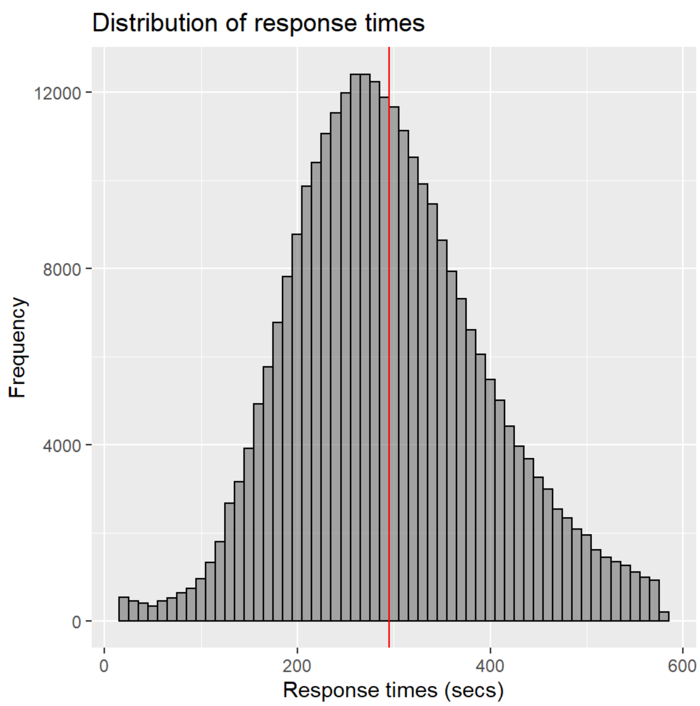
```
## [1] 12.50986
```

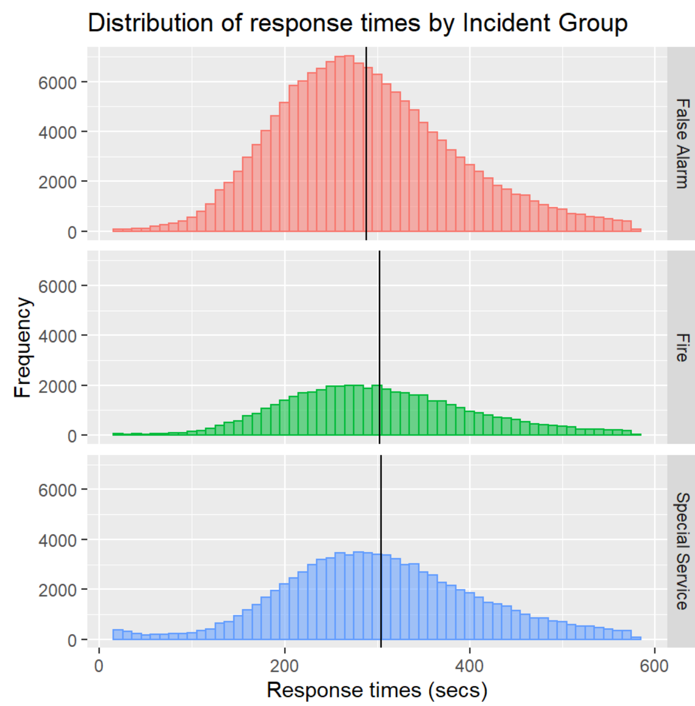
Calculating response time of each incident

```
# Average time by Incident Group
(time_incident <- fire_data_time %>%
  group_by(IncidentGroup) %>%
  summarise(n = n(),avg_time = mean(FirstPumpArriving_AttendanceTime)))
```

```
## # A tibble: 3 x 3
##   IncidentGroup        n avg_time
##   <chr>            <int>    <dbl>
## 1 False Alarm     156635     288.
## 2 Fire             48610     303.
## 3 Special Service  85890     304.
```

```
# Create chart to illustrate distribution of response time
grid.arrange(
  ggplot(fire_data_time, aes(x=FirstPumpArriving_AttendanceTime)) +
    geom_histogram(binwidth = 10, alpha = 0.5, color = "black") +
    geom_vline(mapping = aes(xintercept = mean(FirstPumpArriving_AttendanceTime)), color = "re
d") +
    labs(x="Response times (secs)", y="Frequency", title="Distribution of response times"),
  ggplot(fire_data_time, aes(x=FirstPumpArriving_AttendanceTime,fill = IncidentGroup, color = In
cidentGroup, alpha = 0.5)) +
    geom_histogram(binwidth = 10) +
    facet_grid(IncidentGroup~.) +
    geom_vline(data = time_incident, mapping = aes(xintercept = avg_time)) +
    labs(x="Response times (secs)", y="Frequency", title="Distribution of response times by Inci
dent Group") +
    theme(legend.position = "none"),
nrow = 1, ncol=2)
```





Summary of special service response times

```
# Filter incident groups and group by special service type
(fire_data_special <- fire_data_time %>%
  filter(IncidentGroup == "Special Service") %>%
  group_by(SpecialServiceType) %>%
  summarise(n = n(),
            mean_time = mean(FirstPumpArriving_AttendanceTime, na.rm = TRUE),
            "10th_percentile" = quantile(FirstPumpArriving_AttendanceTime, probs = 0.1, na.rm =
TRUE),
            "90th_percentile" = quantile(FirstPumpArriving_AttendanceTime, probs = 0.9, na.rm =
TRUE) ) %>%
  mutate("%_of_total" = paste(round(n/sum(n)*100, digits = 1),"%"), .after = n) %>%
  arrange(desc(n)))
```

```
## # A tibble: 21 x 6
##    SpecialServiceType              n `%_of_total` mean_time `10th_percentile` `90th_percenti
le`
##    <chr>                       <int> <chr>            <dbl>             <dbl>            <d
bl>
##  1 Effecting entry/exit        22312 26 %              305.              184              4
43
##  2 Flooding                    19405 22.6 %            310.              191              4
46
##  3 RTC                         11064 12.9 %            299.              169              4
48
##  4 No action (not false alarm)  7190 8.4 %             310.              187              4
49
##  5 Lift Release                 4340 5.1 %             293.              177              4
23
##  6 Assist other agencies        4132 4.8 %             309.              188.             4
45
##  7 Making Safe (not RTC)        3124 3.6 %             303.              180              4
44.
##  8 Hazardous Materials incident  2414 2.8 %            305.              187              4
37.
##  9 Animal assistance incidents  2006 2.3 %             321.              192              4
68
## 10 Spills and Leaks (not RTC)   1883 2.2 %             323.              196              4
72.
## # ... with 11 more rows
```

A t-test comparing Ealing and Greenwich

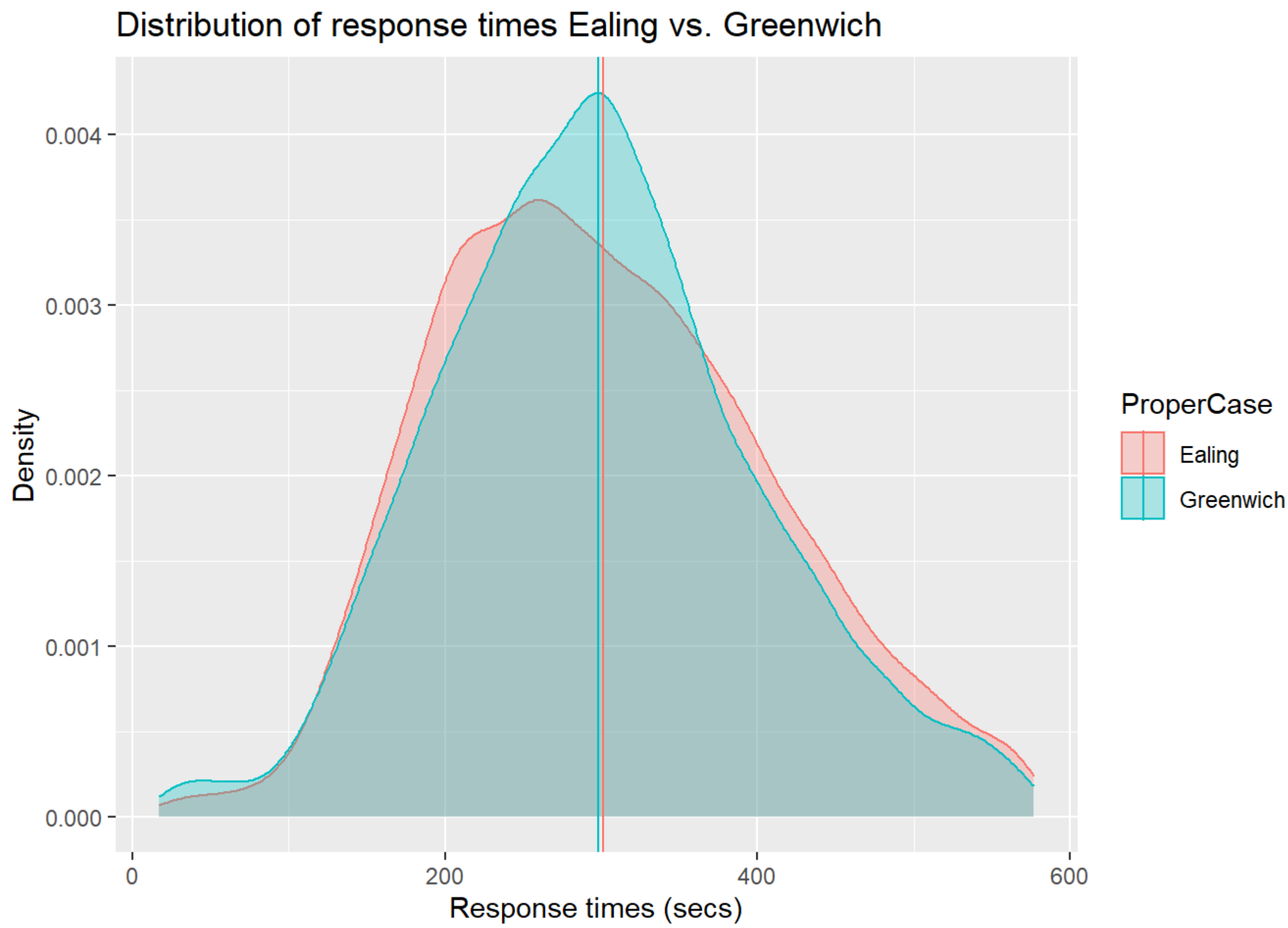
```
# Filter data to contain only Ealing and Greenwich
fire_data_ealing_green <- fire_data_time %>%
  filter(ProperCase == "Ealing" | ProperCase == "Greenwich")

(fire_data_ealing_green.summary <- fire_data_ealing_green %>%
  group_by(ProperCase) %>%
  summarise(n = n(), mean = mean(FirstPumpArriving_AttendanceTime)))
```

```
## # A tibble: 2 x 3
##   ProperCase     n  mean
##   <chr>      <int> <dbl>
## 1 Ealing      9827  301.
## 2 Greenwich   8716  298.
```

```
# Response time distribution of Ealing and Greenwich

ggplot(data = fire_data_ealing_green, aes(x = FirstPumpArriving_AttendanceTime, y = ..density..,
fill = ProperCase, color = ProperCase)) +
  geom_density(alpha = 0.3) +
  geom_vline(data = fire_data_ealing_green.summary, mapping = aes(xintercept = mean, color = Pro
perCase), alpha = 1) +
  labs(x="Response times (secs)", y="Density", title="Distribution of response times Ealing vs.
Greenwich")
```




```
# t-test of responding time shows that mean of Ealing and Greenwich are significantly different
P = 0.037
t.test(FirstPumpArriving_AttendanceTime~ProperCase, data = fire_data_ealing_green)
```

```
##
##  Welch Two Sample t-test
##
## data:  FirstPumpArriving_AttendanceTime by ProperCase
## t = 2.078, df = 18432, p-value = 0.03772
## alternative hypothesis: true difference in means between group Ealing and group Greenwich is
not equal to 0
## 95 percent confidence interval:
##  0.1795543 6.1500834
## sample estimates:
##    mean in group Ealing mean in group Greenwich
##                301.2497                298.0849
```

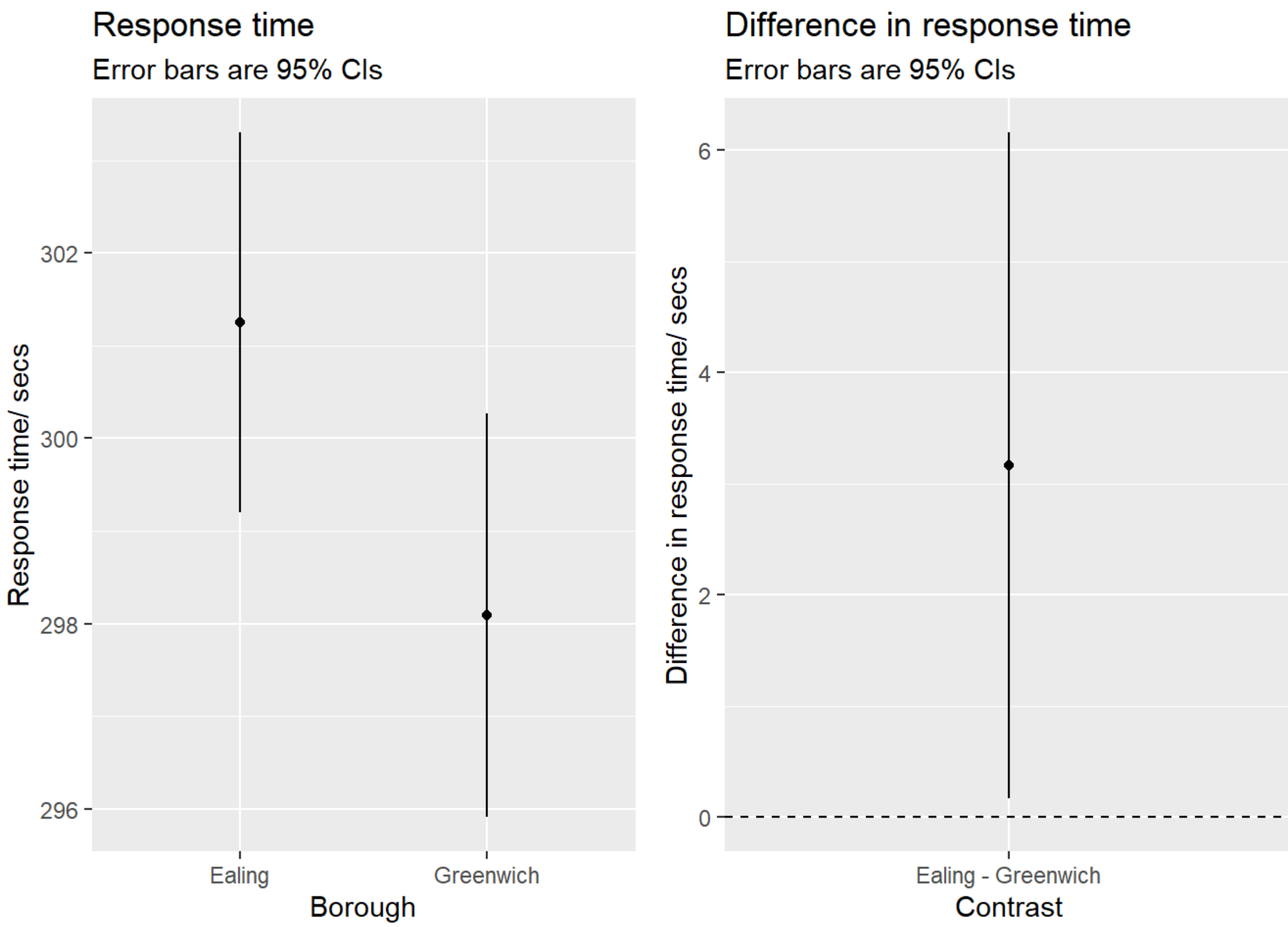
```
# Calculate mean difference and Confident Interval of responding time
m.time.place <- lm(FirstPumpArriving_AttendanceTime~ProperCase, data = fire_data_ealing_green)
(m.time.place.emm <- emmeans(m.time.place, ~ProperCase))
```

```
##  ProperCase emmean   SE    df lower.CL upper.CL
##  Ealing        301 1.05 18541      299      303
##  Greenwich     298 1.11 18541      296      300
##
## Confidence level used: 0.95
```

```
(m.time.place.constrast <- confint(pairs(m.time.place.emm)))
```

```
##  contrast           estimate   SE    df lower.CL upper.CL
##  Ealing - Greenwich     3.16 1.53 18541    0.172     6.16
##
## Confidence level used: 0.95
```

```
# plot the CI range for mean as well as mean difference
grid.arrange(
  ggplot(summary(m.time.place.emm), aes(x = ProperCase, y = emmean, ymin=lower.CL, ymax=upper.C
L)) +
    geom_point() + geom_linerange() +
    labs(y="Response time/ secs", x="Borough", subtitle="Error bars are 95% CIs", title="Respons
e time"),
  ggplot(m.time.place.constrast, aes(x=contrast, y=estimate, ymin=lower.CL,ymax=upper.CL)) +
    geom_point() + geom_linerange() +
    labs(y="Difference in response time/ secs", x="Contrast", subtitle="Error bars are 95% CIs",
title="Difference in response time") +
    geom_hline(yintercept = 0, lty = 2),
nrow=1, ncol=2)
```



Section 2

Data preparation

This report presents the results of the analyses requested by panel of Fire service managers and local politicians. The data use in this report is provided by London Fire Brigade which contains **322,375 incidents** from 2019 - 2022. The analysis in this report will focus primary on cost and response time aspect of the data.

There were multiple outliers for cost in the dataset, which may cause from rare costly events that occur during the analyse period. Data for incidents that cost more than GBP 100k were removed prior to the analyses reported below, leaving **320,244 incidents** for the analysis of cost.

Similarly, there were a large portion of long-tail data for the response time in the dataset, which have been removed by using statistical method for the purpose of this analysis. There were also some data that has no response time which have also been removed. After the data cleaning process, **291,135 incidents** were left for the analysis of response time.

Data analysis

We begin with summaries of Cost of responding to Fire (Table 1).

Table 1: Cost of responding to Fire by type

IncidentGroup	total_cost	avg_cost
False Alarm	61249812	378.38
Fire	39676816	772.43

- Table 1 illustrate the total cost and average cost of fire and false alarm. Comparing between the two, total cost of false alarm is twice as much as fire.
- In contrast, average cost per incident of fire is significantly higher than false alarm, which could be explain by the concentration of rare costly incidents which can be observed in the fire incident.

Objective of the next section below is to examine the response time of incidents during the analysis period.

Figure 1: Distribution of response time

- Figure 1 highlights the distribution of response times. The left-hand side (LHS) chart illustrate that the distribution of response is normally distributed. Similarly, the right-hand side (RHS) chart also echoing the same pattern that response time is normally distributed in all Incident Group.
- Focusing on the RHS chart, false alarm has the highest frequency among all incident group follow by special services and fire, respectively. While there is no noticeable difference in average response time between special services and fire, false alarm incident is having approximately 20 seconds faster response time comparing to the rest of the incident.

Table 2: Special services case and response time

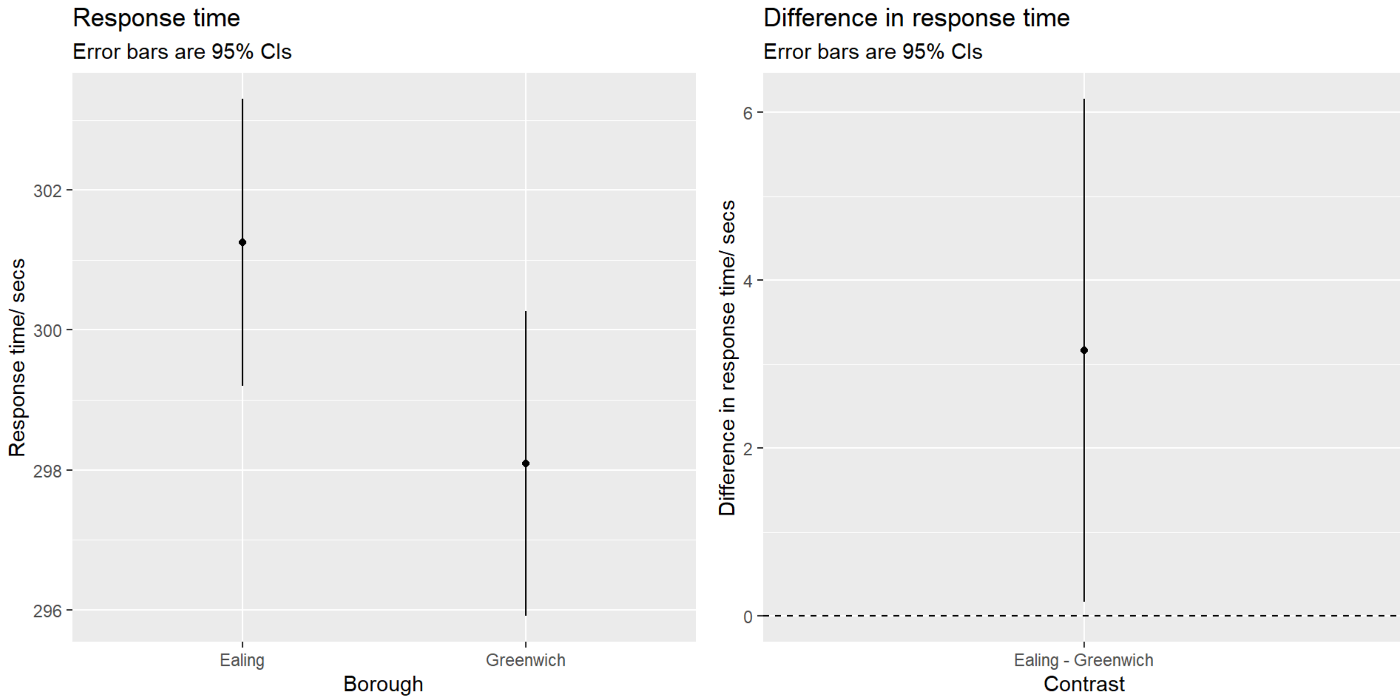
SpecialServiceType	n	%_of_total	mean_time	10th_percentile	90th_percentile
Effecting entry/exit	22312	26 %	304.88	184.0	443.0
Flooding	19405	22.6 %	310.40	191.0	446.0
RTC	11064	12.9 %	298.92	169.0	448.0
No action (not false alarm)	7190	8.4 %	309.52	187.0	449.0

SpecialServiceType	n	%_of_total	mean_time	10th_percentile	90th_percentile
Lift Release	4340	5.1 %	292.51	177.0	423.0
Assist other agencies	4132	4.8 %	309.15	188.1	445.0
Making Safe (not RTC)	3124	3.6 %	302.77	180.0	443.7
Hazardous Materials incident	2414	2.8 %	305.12	187.0	436.7
Animal assistance incidents	2006	2.3 %	321.43	192.0	468.0
Spills and Leaks (not RTC)	1883	2.2 %	323.49	196.0	471.8
Advice Only	1748	2 %	309.81	189.7	440.3
Medical Incident	1655	1.9 %	247.62	37.4	421.0
Other rescue/release of persons	1149	1.3 %	314.84	189.0	456.0
Removal of objects from people	1100	1.3 %	252.21	29.0	449.1
Other Transport incident	784	0.9 %	297.70	159.3	443.0
Suicide/attempts	664	0.8 %	306.61	183.3	451.0
Evacuation (no fire)	626	0.7 %	309.15	190.0	445.5
Rescue or evacuation from water	150	0.2 %	295.79	174.9	422.2
Stand By	142	0.2 %	304.37	176.0	435.9
Water provision	1	0 %	245.00	245.0	245.0
NA	1	0 %	169.00	169.0	169.0

- As requested by the panel, Table 3 outline the type of special services performed during the analysis period sorted by the frequency in the descending order.
- The top 5 most common occurrence are Effecting entry/exit, Flooding, RTC, No action (not false alarm) and Lift Release respectively, which represented 75% of all occurrence.
- The average response time is approximately the same for majority of special service types. In term of the response time range, Medical Incident and Removal of objects from people are having a noticeably wider range which could be due to the severity of the incident.

A t-test comparing Ealing and Greenwich

As per the panel’s request, below chart illustrate the comparison of response time between Ealing and Greenwich.



- The t-test shows that Greenwich's response time is significantly less than that of Ealing $t(18432) = 2.078, p = 0.03772$
- The mean in Ealing's response time is 301.25 seconds 95% CI [299–303] while the mean in Greenwich's response time is 298.09 seconds 95% CI [296–300] as illustrated in the left-hand side chart. The response time is 3.16 seconds 95% CI [0.17–6.16] smaller at Greenwich compared to Ealing as per the right-hand side chart.

In conclusion, the analysis shows the insight on the response time between two locations, however, the analysis also come with a significant caveat. While the IQR method of outlier detection was applied to the dataset to create normally distributed data, the actual data is not appropriate for using the t-test as the data is considered to be positively skewed. This may mislead the interpretation of the t-test result.

In addition, IQR method may excluded some valuable data which has extreme value and fail to properly capture the whole picture. We recommend the panel to initiate further analyses that are more suitable for the data.