Stats15 F24
Dec 6th, 2024
Professor Gould
Team: CARE (Group D)
Members: Emma Teng, Ashlyn Steen, Colin Sohn, Raka Sudijono

Stats 15 Final Project Report

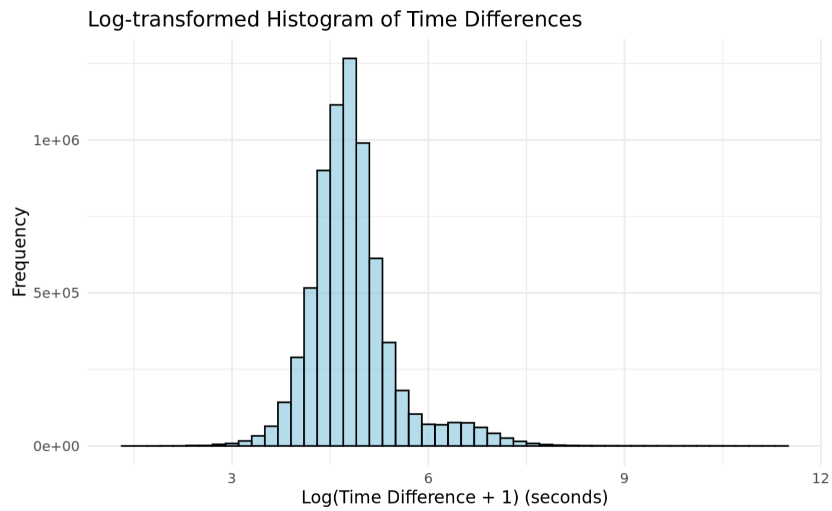
Abstract:
In this study, we worked to uncover the most important factors that affect the response time of the LA Fire Department. Here, response time is calculated as the time between the incident creation time, when the incident is created, and dispatch time, when teams are dispatched to respond to the incident. We concluded the factors with the most association are: time of day, PPE Level (EMS vs Non-EMS), and Unit Type (Engine, truck, etc.). We also uncovered a few other interesting associations, such as the relationship between dispatch status and how dispersed the response teams are based on the time of day. Using a random forest model, we confirmed that the best variables for predicting response time are dispatch sequence and emergency code. Finally, we found the combination of variables that lead to the fastest response time.

Introduction:
The data we used for this study, "LAFD Response Metrics", is collected and maintained by the Los Angeles Fire Department. It was created in 2018 and last updated in July, 2024. The dataset includes 7.2 million rows/observations, each representing a fire department response, and 11 columns containing variables that describe the response. The data is automatically collected by the LAFD Computer Aided Dispatch (CAD) system that records time stamps based on triggered events.

The main question we wanted to answer with this study is which factors affect LAFD response times. We were interested in looking at the association between response time and numerous variables, including PPE, Dispatch Status, Unit Type, etc. Along the same line of inquiry, we were also interested in determining which variables would be most useful in predicting response time. Further, we planned to consider how outside factors, such as the time of day an incident occurred, would also affect response time. Once equipped with these answers, we aimed to piece together the optimal set of conditions to create the fastest response time possible.

Data Cleaning/Data Exploration:
To answer our questions, we first had to figure out some relationships and distributions of different variables. We started by trying to figure out the distribution of response times. However, since a response time variable did not exist, we had to create one called route_response using the data we were given. We subtracted incident creation time from dispatch time for this

variable. Then, we discovered that there were some negative response times. We also found the mean response time, but it was negative which shouldn't be physically possible. Since incident creation time and dispatch time only take into account the time of day (and ignore the date), we discovered that there were major outliers when an incident was created shortly before midnight and the dispatch time occurred the next day. Since the data set doesn't record the actual date, just the time of day, it made the difference between the two nearly -24 hours, pulling the mean difference in data negative. To fix this, we added 24 hours to the dispatch time of incidents in which the response time was negative. Calculating the mean afterwards, we got an average response time of 185 seconds. To see the distribution we made the following histogram.



This histogram has the log response time on the x axis (we used log of the response + 1 so that the data varied less in our visualization so it would be easier to see) and the frequency on the y axis. Log in R refers to natural log. Using this knowledge, the response time peaks at around 148 seconds and is slightly right skewed.

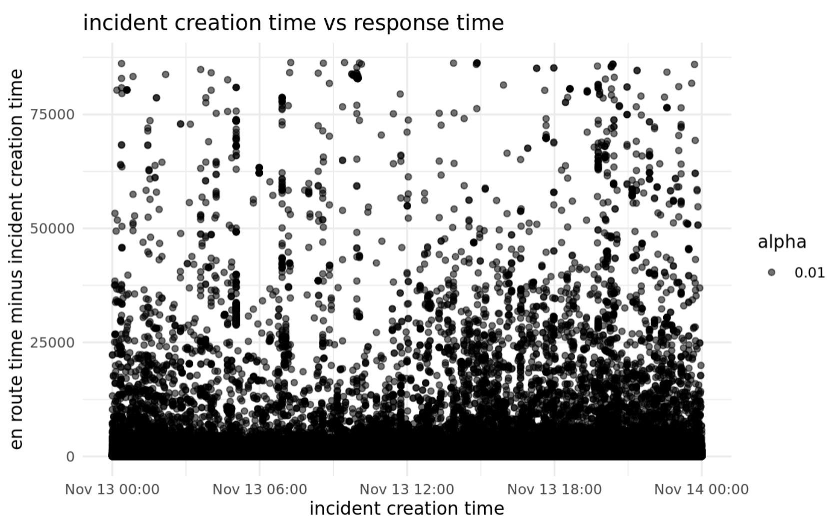
Further, we discovered that there were a number of incidents where the recorded incident creation time occurred after the dispatch time, which shouldn't be physically possible since the incident call needs to occur before an alert is signaled. This prompted us to look at the data collection methods for the data set, where we discovered that dispatch time is automatically recorded by the CAD system, but incident creation time is manually logged by the responder, and therefore subject to human error if it is recorded too late. Removing the incidents where this error occurred, we could proceed with our study.

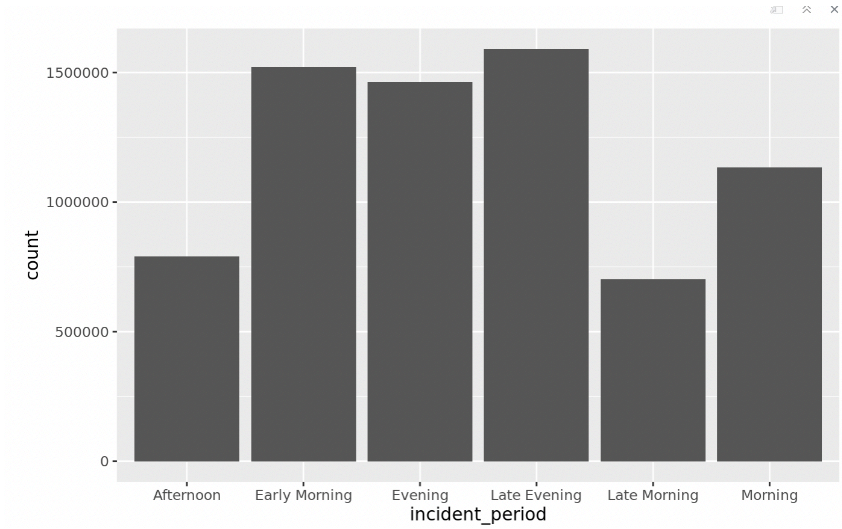
When exploring our data, we also looked at the largest dispatch sequence numbers. There was one incident that mandated 166 vehicles to be dispatched from first in district 23. This was a strange phenomenon that we attempted to look into. Out of the 50 largest dispatch sequences in 42 came from District 23. This occurred during the second quarter of the year, a period that did not witness any significant local fires, except for the Caldor Fire in Northern California. While the Caldor Fire did trigger a statewide response, including mutual aid from various districts, it didn't explain why District 23 was the only district with such unusually high dispatch numbers. The reasons behind the high dispatch sequences from District 23 remain unclear, leading to a dead-end in our research as no other major fires or events during this period directly correlate with the surge in activity at Station 23.

Data Analysis Methods:
To answer the question “What variables are the most strongly associated with LA Fire Department response times?”, we used a combination of visualizations, comparing means and hypothesis testing. To answer the question “What variables would be most useful in predicting LAFD response times?”, we used a random forest with a portion of the data and created a variable importance plot. To figure out “What combination of variables results in the quickest response time?”, we calculated variable averages and compared them.

Data Analysis Results:
Our first question to explore was “What variables are the most strongly associated with LA Fire Department response times?”.

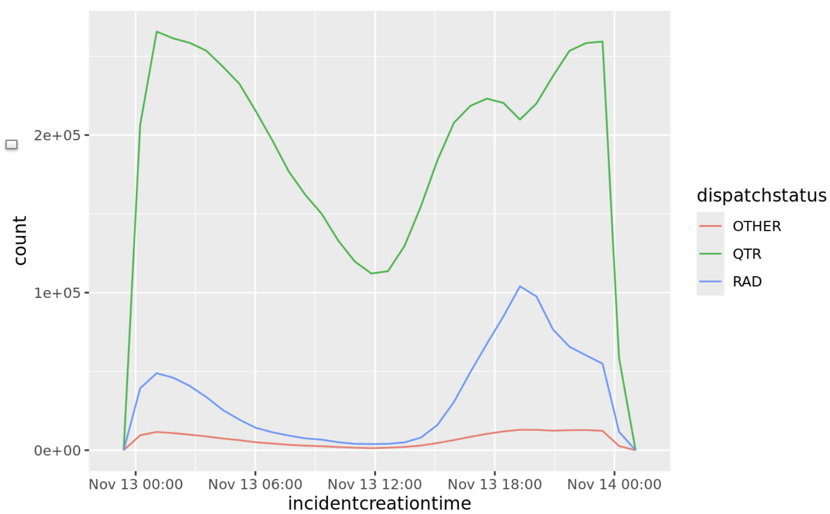
We found through data visualizations that time of day seemed to have a relationship with LAFD response times.

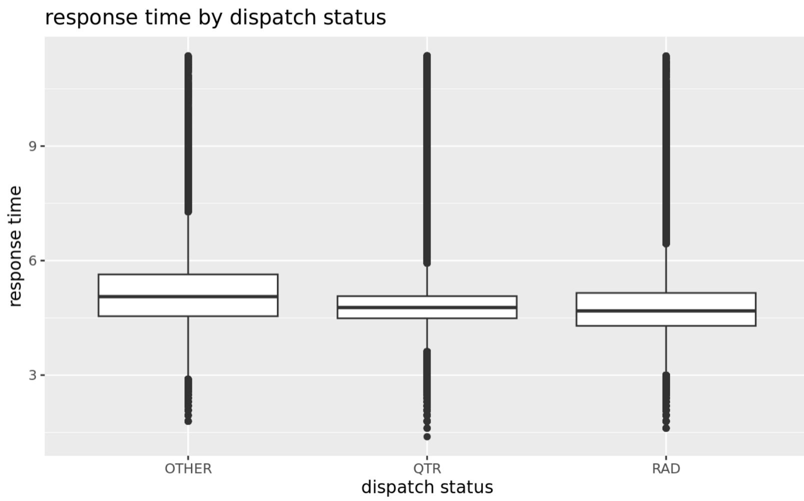




We found that there seemed to be a smaller response time in the late morning, so we investigated further to find that there also seemed to be less incidents during this time of day, as shown in the graph above. Therefore, we concluded that perhaps the time of day and the frequency of incidents play a role in determining the response time.

Next, we looked at the relationship between dispatch status and response time.

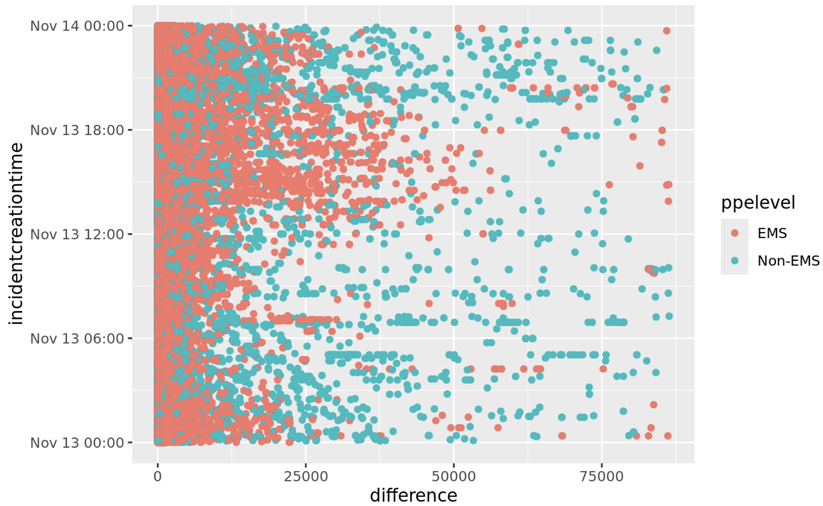




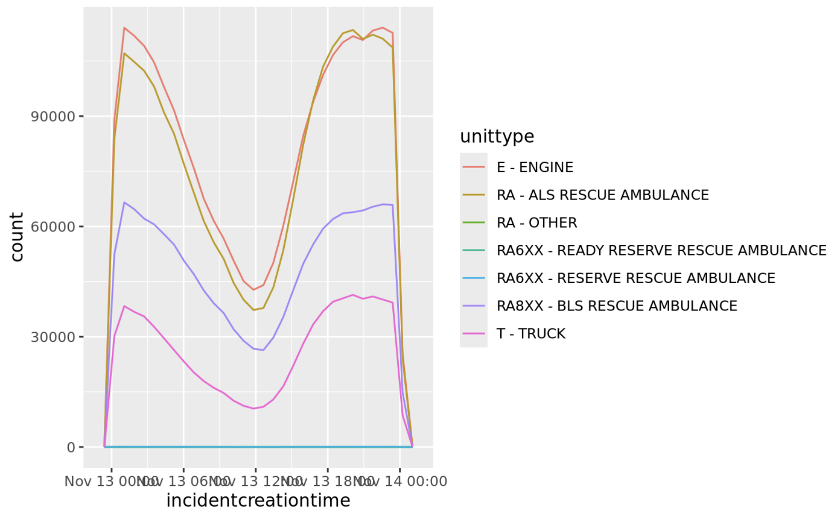
We looked at dispatch status to see if there was an association with response time depending on the responding units status at the time of dispatch: “QTR” when the unit responded from headquarters, “RAD” when they responded from radio, and “OTHER” for everywhere else. Creating a boxplot, we found that the median response times of different dispatch status weren't much different from one another. Looking at the frequency polygraph however, it is interesting to note that RAD incidents increase in frequency during the most frequent incident times. This would make sense because the units are outside of quarters more often during peak hours, leading to more radio responses.

We decided to investigate further and answer “Is there a difference in response times between QTR and RAD incidents?” To do this, we carried out a hypothesis test. Our null hypothesis claimed that there is no difference in the LAFD response times for QTR vs RAD incidents. Our alternative hypothesis was that QTR incidents have a faster response time than RAD incidents. This can also be stated as the absolute difference between RAD incidents and QTR incidents is greater than 0. Our test statistic was the difference in mean response time for RAD and QTR, which was 32 seconds (147-115). We decided to use a significance level of .05. We got a p value of 2.2 * 10^-16. This value is much less than our significance value of .05, so we rejected our null hypothesis. We can conclude that getting a difference between RAD and QTR response times this large is unlikely to occur by chance, so the true difference in RAD and QTR average response times are likely to be greater than 0 in the long run. We also decided to try bootstrapping to test the same question. We got a higher p value of .06. Using this p value, we would fail to reject the null hypothesis. With conflicting results, we decide to trust the bootstrapping method more and be more cautious. Therefore, we did not reject the null hypothesis that there is no difference in response times between QTR and RAD incidents.

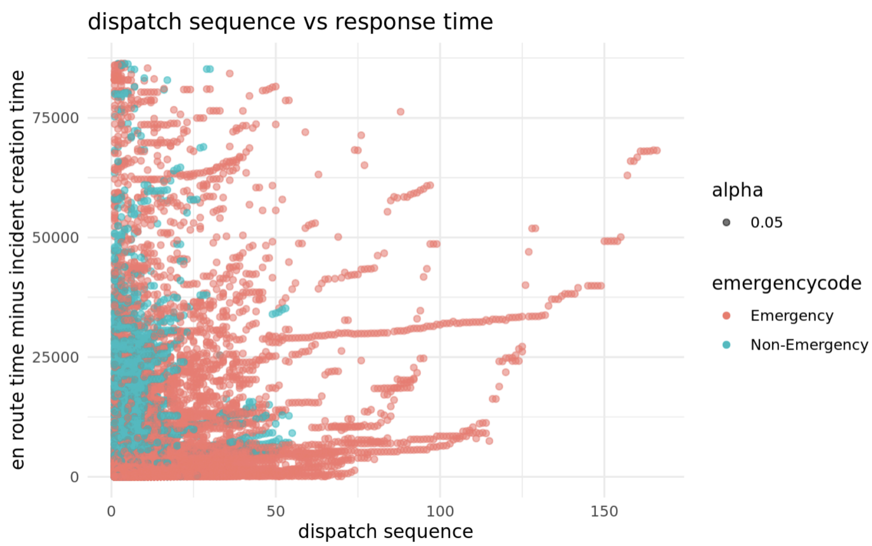
We followed by trying to figure out if there was a difference in response time for different PPE levels.



There seems to be a quicker response time for EMS than Non-EMS. EMS incidents require a minimal amount of medical equipment and are therefore required to have a smaller response time than Non-EMS incidents which require full medical equipment. Our graph demonstrates this! This led us to calculate the mean response time of EMS and Non-EMS. The mean EMS response time was 177.1314 seconds, while the mean Non-EMS time was 229.3104. This drastic difference led us to carry out a significance test between them later. The question we sought to answer was "Is there a difference in response times between EMS and Non-EMS incidents?" To do this, we carried out a hypothesis test. Our null hypothesis claimed that there is no difference in the LAFD response times for EMS vs Non-EMS incidents. Our alternative hypothesis was that EMS incidents have a smaller response time than Non-EMS incidents. This can also be stated as the difference between Non-EMS incidents and EMS incidents is greater than 0. Our test statistic was 21.9 seconds (143.9-122). Using a t-test, we got a P value of roughly .05, and given that it borders with our significance level, supported by our visual analysis of the dotplot, we decided to reject our null hypothesis. We can conclude that getting a difference between Non-EMS and EMS response times this large is unlikely to occur by chance, so the true difference in Non-EMS and EMS average response times are likely to be greater than 0 in the long run.

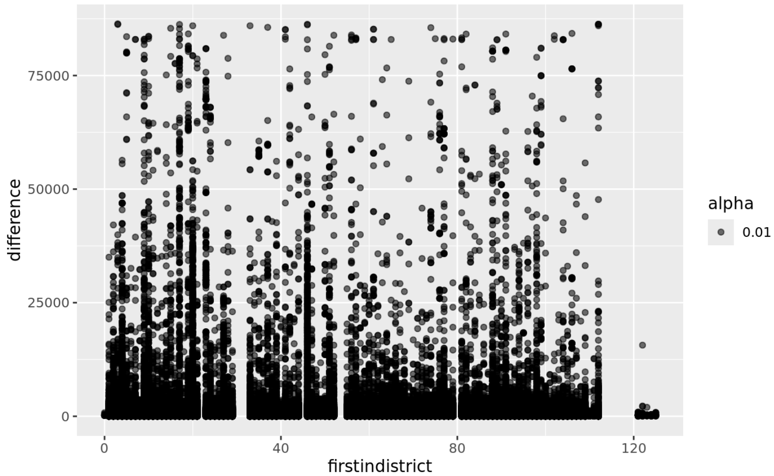


We found that there seemed to be different average response times depending on the unit type. The quickest response time (in seconds) was engines (164), followed by trucks (179), ALS rescue ambulances (197), and BLS rescue ambulances (207). This did not seem to be tied to the frequency that each unit type was used, as shown in the graph. Next, we looked for a relationship between dispatch sequence, emergency code, and response time.

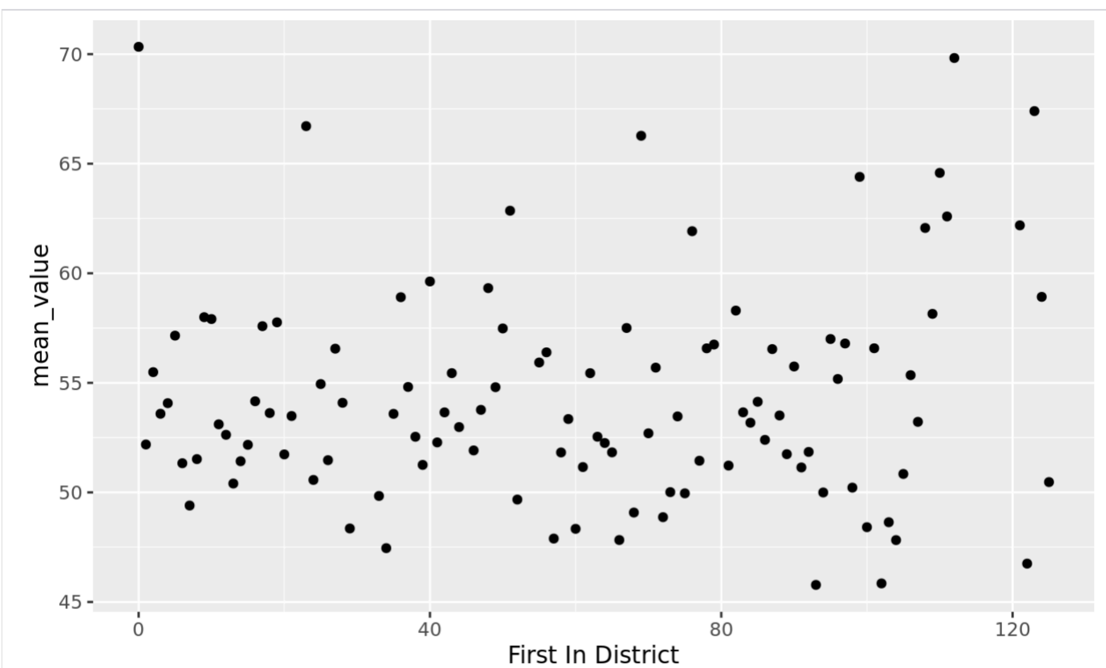


We were curious to see the response time depending on the dispatch sequence number (the number in order a unit is dispatched to a particular incident.) There didn't seem to be much of an association between dispatch sequence and response time, however, only emergencies seemed to have very long dispatch sequences. This makes sense because there probably wouldn't need to be as many vehicles for a non-emergency! Interestingly, there seems to have been one especially severe incident that warranted 166 units to be dispatched from first in district 23. Unfortunately, we were unable to discover any more information about this incident and what made it so unique. This Station is located in Pacific Palisades which is near Will Rogers State Beach!

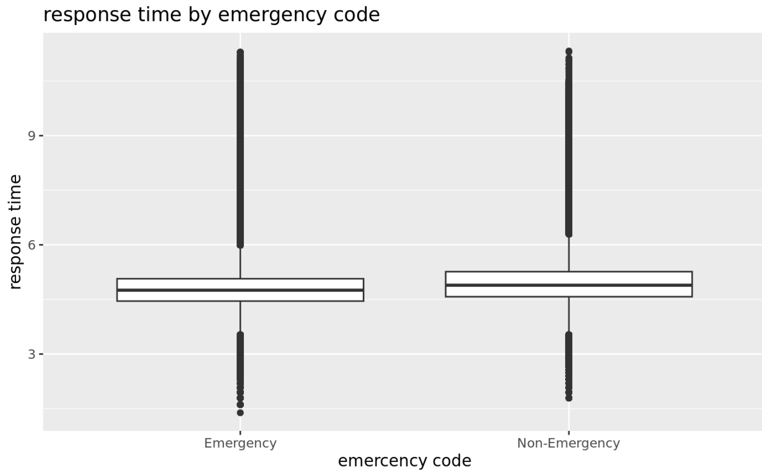
Finally, we looked at firstindistrict and tried to find any patterns in response times.



After the anomaly of District 23, we wanted to see if there was a difference in response time for each of the districts. This was difficult to tell due to the massive amount of data and large number of districts. There didn't seem to be any clear correlations based on our visualization. To confirm this finding, we grouped the data by each district and found the mean response time for each. The resulting visualization confirmed our hypothesis that there was no clear correlation between district and response time as the points were scattered with no strong association. Certain districts may have faster or slower response times, but that can likely be attributed to chance and conditions in the specific incidents happening in their district.

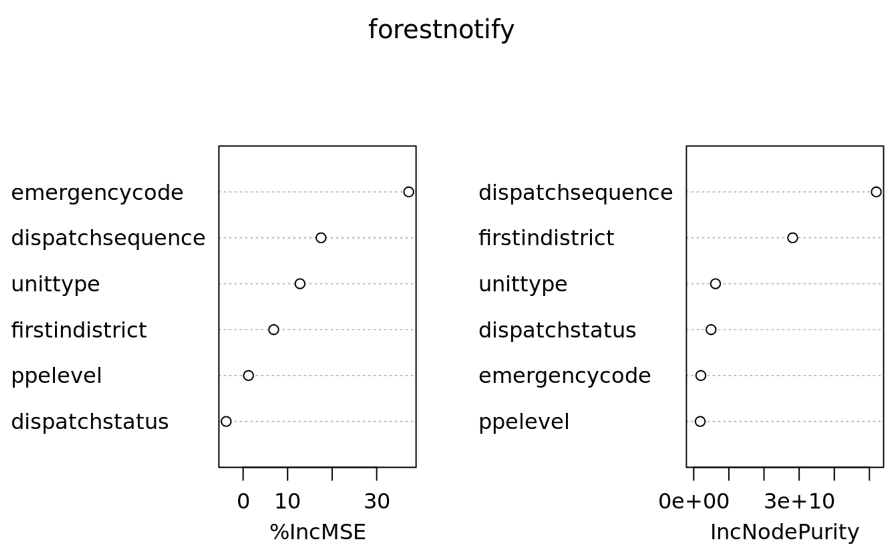


We decided next to look at the emergencycode variable.



The distributions of emergency response times vs non-emergency response times don't seem to differ very much. Emergency incidents have a lower average response time and they seem to vary less, but there is not an extreme difference between the two. The NFPA requires that fire departments respond to emergencies within 4 minutes, whereas non-emergencies do not have the same time constraints. The combination of all of these visualizations and calculations led us to conclude that a fast response time is positively associated with late morning, EMS incidents, and engines.

The next question we investigated was "What variables would be most useful in predicting LAFD response times?". To answer this question, we fit a random forest regression model on a random sample of our data (using the entire dataset would have exceeded R's computing limit). Next, we generated an importance plot to see which factors contributed the most to determining response time.



Using this plot, it seems that emergency code and dispatch sequence are the best predictors for response time. This is followed by first in district and unit type. This conclusion supports our earlier analysis of associations between these variables and response time. It is important to note here, however, that even if a variable such as PPE level doesn't appear to be the best predictor for response time in this model, it doesn't mean that it contributes less to determining the length of the response time overall.

Lastly, we wanted to answer "What combination of variables results in the quickest response time?" We first discovered that firstindistrict 102, emergencycode "Emergency", dispatchsequence 1, dispatchstatus "QTR", unittype "Engine", and ppelevel "EMS", had the quickest response times in their respective categories. When we looked at all of the incidents that had al1 of these characteristics, their average response time was 105 seconds. There were 7490 of these incidents. This is much lower than the 175 second average response time. Each of the fastest variables in their categories had a quicker average response time than the total average, spanning from 122-172 seconds. However, when combining them all, they had a very quick response time of 105 seconds.

Summary:

In this study, we discovered that the variables most associated with response time are time of day, PPE level, emergency code and unit type. While time of day is an outside variable that determines uncontrollable factors such as traffic levels, the remaining 3 are all processes that happen within the fire station. Similarly, using a random forest model to predict response time, we concluded that dispatch sequence and emergency code were the best predictors for response time. Finally, we found that the combination of variables that result in the quickest response time are firstindistrict 102, emergencycode "Emergency", dispatchsequence 1, dispatchstatus "QTR", unittype "Engine", and ppelevel "EMS". These findings are consistent with our conclusions from earlier graphs.

Using these findings, LAFD could evaluate how they can improve their operations to shorten response times. For example, since the "engine" unit type has a faster response time than "truck", we can consider how to make "truck" units respond faster through better preparation or having a quicker turnaround time for the incidents that require them. Another example could be to study district 102's operations and examine why they have a faster average response times than all other districts. While this could be attributed to random chance in the incidents that occurred there, it could also be due to district 102's unique features that minimize their response time.

Some recommended next steps from us include researching how much of a role each of the associated variables mentioned above play in affecting response time. Now that we have found they are associated with response time, we can attempt to find the weight of each to determine which variable is the most important and should be addressed by LAFD first with regards to improving response efficiency. We could also launch a deeper examination into large incidents that require more response units than usual and analyze their respective response times, emergency levels, etc. to improve efficiency in these cases specifically.

Fire Department response times are a critical measure that can mean life or death for those hanging in the balance. It can also mean the preservation or destruction of a home, a business, or a community. Studying fire department response times is an important task that can help units better allocate resources, improve response efficiency and ultimately result in better outcomes for the firefighters risking their lives and the residents depending on them.