Assignment 1 - Initial Project Plan

Student: Natasha Sharma

Date Submitted: 17th Sep 2023

Task Selected: "Response Time"

Task Description:

As an analyst we are supposed to predict the time taken for an Ambulance to respond to a call based on the data provided. We will have to use one of the Machine Learning algorithms to complete this task. We will also have to analyze the data to figure out what factors will affect the response time and how those factors can be used in our ML algorithm.

Final Predictive Value:

We are trying to prexsdict the time taken in ***seconds*** for EMS to

1. Respond
2. Dispatch
3. Reach

the scene of an incident. This will be for different,

- Boroughs where the incident happened (location)
- Time, Day of week and month of the incident
- Weather conditions at the time the incident was reported

Factors Considering:

- Day of the week (M, T, W, Th, F, Sa, Sn)
- Borough where the incident was reported
- Zip code where the incident was reported
- Time of day
- Incident Dispatch Area
- Rain
- Snow
- Heavy Fog
- Number of incidents on that day

Data Sources Considering:

- NYC FDNY Emergency Medical Services Ambulance Calls, 2008-2016
- NYC Weather Data 2008-2016

Next Steps:

- Clean the data.
    - Determine which columns are needed for this model
    - Remove any rows which have missing data or null values
- Determine which predictors might have correlation.
- Merge the two different datasets.
- Divide the dataset into two parts “training data” and “testing data”
- Determine if clustering is required for this data
    - Choose features from the data that can be used to create clusters
    - Apply a clustering algorithm to create clusters
- Determine which algorithm will be best to predict the response time.

Data Generating Process / Bias Template

Student: ***Natasha Sharma***

UAlbany Id - 001573201

Date Submitted: *8th October 2023*

Task Selected: *Response Time*

Before we get to working with the data and building a predictive model, in this document you will provide an assessment of what you, the model builder, perceive to be the data generating process for your data sources and potential sources of bias that could result from either the data or the model building process.

Data Generating Process

Since you are role-playing a data analyst responding to your supervisor's request to build a predictive model, you will be working with data that you didn't have a hand in originally collecting, and you may not fully understand. This is common in many practices as a data analyst, and you have to build a healthy skepticism of **any** dataset you work with, even if you created it yourself. Before you can build your predictive model, you have to first understand how the data was generated, and what potential sources of bias might be present that could skew your results. This first section will focus on the data and its potential sources of bias; the next section will focus on issues that could arise with the modeling process itself.

Item	*Ambulance Calls for Service*	*Training Data EMS calls*	*NYC Weather Data 2008-2016 by Date*
Dates of Coverage	{Please identify the specific dates that will be used for this dataset in your modeling effort}	2008-01-01 00:00:01 UTC till 2016-9-9 23:59:44 UTC	2008-01-01 2016-09-09
Frequency of data collection	{how often is the data collected? After every incident? Daily? Yearly?}	Data is collected after every incident.	Data is collected on daily basis. Since, there is only one row per date, the data points reflect the value for the whole day.
Agency / Organization collecting the data	{who specifically is collecting the data? Please avoid using general references like "government" or "police}	Data is being collected based upon 911 calls. So, it looks like it is being collected by the EMS providers with help of PCR report.	The data may have been collected by National Weather Service.

Original Unit of Analysis	{What is the original unit of analysis for the data as provided? Calls for service? Census tracts? Cities?}	Timestamp (Date, hour, minutes and seconds) of the Incident call.	Speed – mi/h Precipitation – in Temperature – F Snowfall - in
Transformed Unit of Analysis	{i.e. are you modifying the call data to support your model? Hint: if you are doing "demand" model you will be aggregating the data.}	We will be taking the difference between the different timestamp columns (**incident_dt, first_act_dt, first_on_scene_dt etc.**). So the transformed unit will be in seconds.	We will be normalizing the data. However, the units of the resultant data will be same.
Data Generation Description	{here, I want you in your own words to describe how you think the data was generated. Think 2-3 sentences.}	Once an incident is reported, the date and time of the incident is noted. Severity codes are then assigned by 911 call dispatcher. Timestamps are then reported for each part of the process of dispatching an Ambulance. EMS responders arrive at the scene and note the rest of the timestamps related to arrival at scene and first to hospital.	Weather stations collect meteorological data using their equipment. There should be different methods of measuring each data point. For example – Temperature is measured by thermometer. Rain gauge measures rainfall etc.
Data Collector	{Who collects the data? A dispatcher? A Census taker?}	Data collection is done by dispatchers and EMS providers who handle the incidents.	People working in Weather Stations collect this data and then share it with National Weather Service.
Triggering Process	{What triggers the data to be collected? A call for service? A yearly survey process?}	The data is collected whenever a call for service is made.	Weather Conditions trigger the data to be collected, so that right precautions can be taken for public safety.
Process Alignment	{What system captures the data? Is it hand entered? What existing business process does the data align with [i.e. data on ATM transactions come from someone using an ATM; data from calls for service come from	The data is being captured via dispatch records for calls coming in to 911.In addition to this, it is also being captured by the ambulance service provider.	Instruments capture the data. There are special instruments at weather stations that help capturing this data.

	dispatch records for calls coming in to 911, etc.]?}		

Potential Sources of Bias

In this project, you are using ambulance calls for service data from New York City during the years 2008-2016 to build your models. While the data is administratively collected from Computer-Aided Dispatch records, it doesn't mean that it is free from differences in data scope or collection that could bias its results from one or more subgroups. In addition, any additional datasets you plan on bringing in also suffer from potential biases that need to be acknowledged at the outset. In this section, for the datasets mentioned above, we will consider some key questions to help flesh out those potential biases.

Item	*Ambulance Calls for Service*	*Training Data EMS calls*	*NYC Weather Data 2008-2016 by Date*
Representativeness	{What is this dataset attempting to capture data about? Is the dataset designed to be representative of the underlying phenomenon that it is attempting to cover?}	The dataset is attempting to record the timestamp of every incident that happens on a particular day and time and how it is being processed from a phone call of an incident to an ambulance arrival to hospital arrival. It is also capturing other information regarding an incident like severity, location etc.	It is attempting to capture the weather condition in NY City Central Park, NY US. Since, the measurements have been taken by special instruments, it should be accurate enough to represent the weather conditions.
Geographic Coverage	{Does this dataset cover all geographic areas of interest for your model? If not, what is missing? Why is it missing?}	The dataset is covering all the Boroughs in New York City. Hence, it looks like it is not missing anything.	It only covers geographical area of NY City Central Park, NY US. Since Central Park comes under Manhattan Borough, we would have to assume that the weather conditions are same for all other boroughs (Brooklyn, Bronx, Queens, Staten Island).

Demographic Coverage	{Does this dataset cover all demographics of interest for your model? Are there individuals of a specific age / race / ethnicity / ability that would not be included in this dataset? If so, what is missing? Why is it missing?}	This dataset does not cover any demographic information. It is missing because the dataset providers want to protect the privacy of individuals involved in the incidents.	Since, It is weather data, demographic coverage does not apply to it.
Temporal Coverage	{Does this dataset cover all time periods of interest for your model? Are specific times, days, months, weeks, years missing from the data effort? If so, what is missing? Why is it missing?}	Yes, it is covering all time periods of interest for our model.	The weather data generalizes conditions for the whole day, whereas the incident calls have specific time during the day. Hence, we would have to assume that the weather conditions apply on all incidents during the day. This dataset is also missing data about the last three months of the year 2016.
Comprehensiveness	{Does the dataset capture all of the relevant features about your subject of interest that, you think, would be relevant for building your model? If not, what is missing and why is it missing? What limitations will those missing features have on the model you want to build?}	The dataset captures all the relevant features about our subject of interest that would be relevant for building our model.	The dataset is relevant except for the fact that it only covers one specific geographical area (Central Park). Also, there are so many N/A and 0 values, which could affect the prediction accuracy and add bias to the model.
System Drift	{Based on your review of the data source, have you identified any specific factors that suggest changes in the systems collecting the data you are using. This could be a result to a changed research design, the inclusion of different questions, or can be seen in variables that have two options for the same	After reviewing the dataset, all the variables seem to be the factors that can help in building the model or could affect the model. Other than that, there aren't any specific factors that could cause system drift.	Since the factor "Name" only covers one geographical location, it hinders the ability to determine the effect of a particular weather condition on an incident that happens outside of that geographical location. Also the factor "TAVG" and "TSUN" are mostly N/A, which

	answer (think “Y” and “True”, “N” and “False”). If present, which variables are these, and what is your plan for addressing these variables? Do you think the potential system drift would dramatically change your underlying model assumptions? Why or why not?		makes them irrelevant to be in the dataset.

Assignment 3

Descriptive Analysis

Student: *Natasha Sharma*

Due Date: *11/05/2023*

Date Submitted: *10/30/2023*

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.

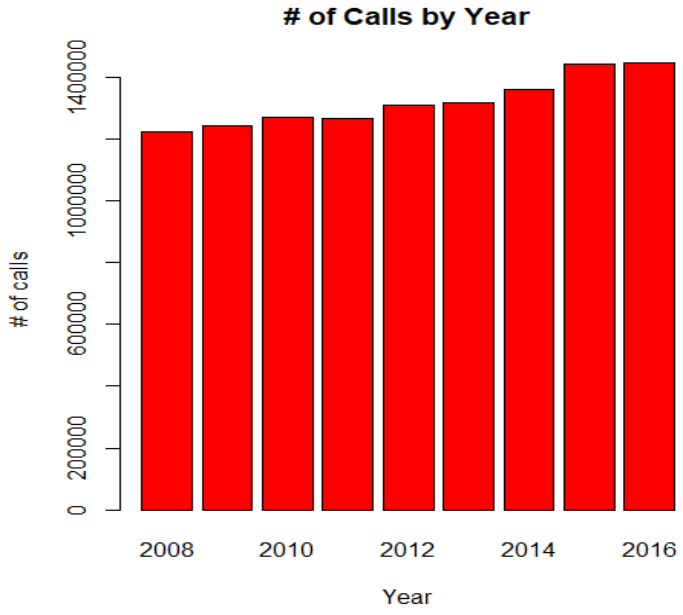
WARNING: This document will become very long (20 or so pages), primarily because you will be posting tables, charts, and then your own initial written assessments of what patterns you are seeing in the data. Please make sure you complete this assignment fully, as it will dramatically help your predictive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

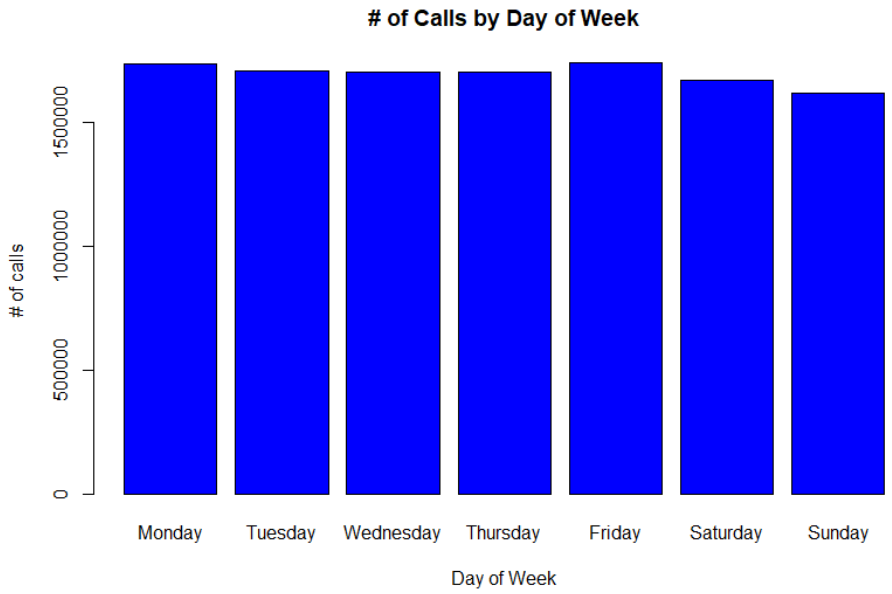
For this section, you will need to complete the following tables, charts, and assessments. There is a space on each page for these items, and you will need to provide not only the formatted items, but also a short (2-3 sentences) description of what patterns you are seeing. Please conduct the requested 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.



What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

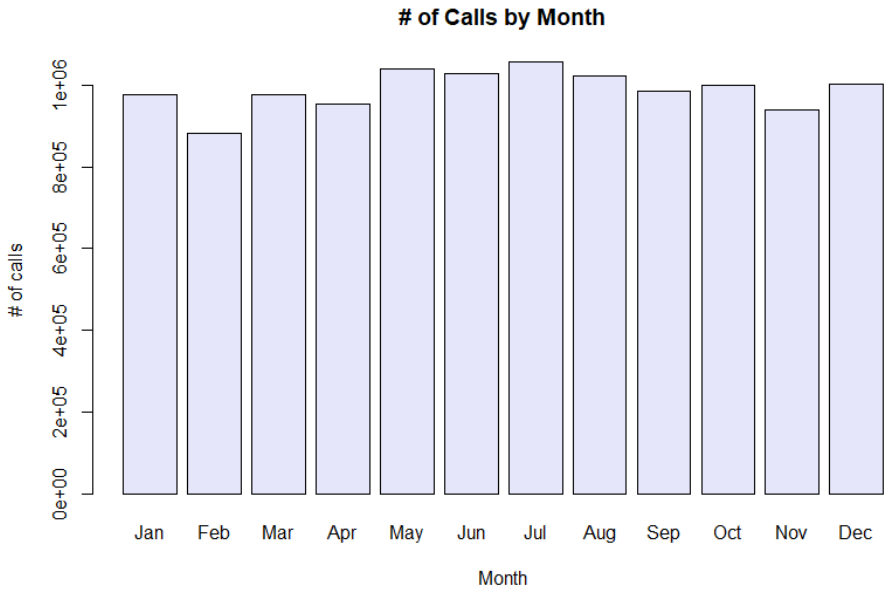
The bar graph of Number of Calls by year shows a trend of increasing the calls every year as we go from 2008 to 2016. The one unusual pattern is that the number of calls is increasing from 2008 to 2010, but there is a slight decrement of calls in 2011. However, the trend is continuing with the increase in the number of calls from 2012 to 2016. Also, there is a much significant increase from 2014 to 2016 when compared to the previous years.



What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The above bar chart representing the number of calls by Day of Week shows that the Number of calls received on Friday are the highest, following by Monday. The graph above also shows that the number of calls on weekends is lowest when compared to other days of the week, especially on Sunday.

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

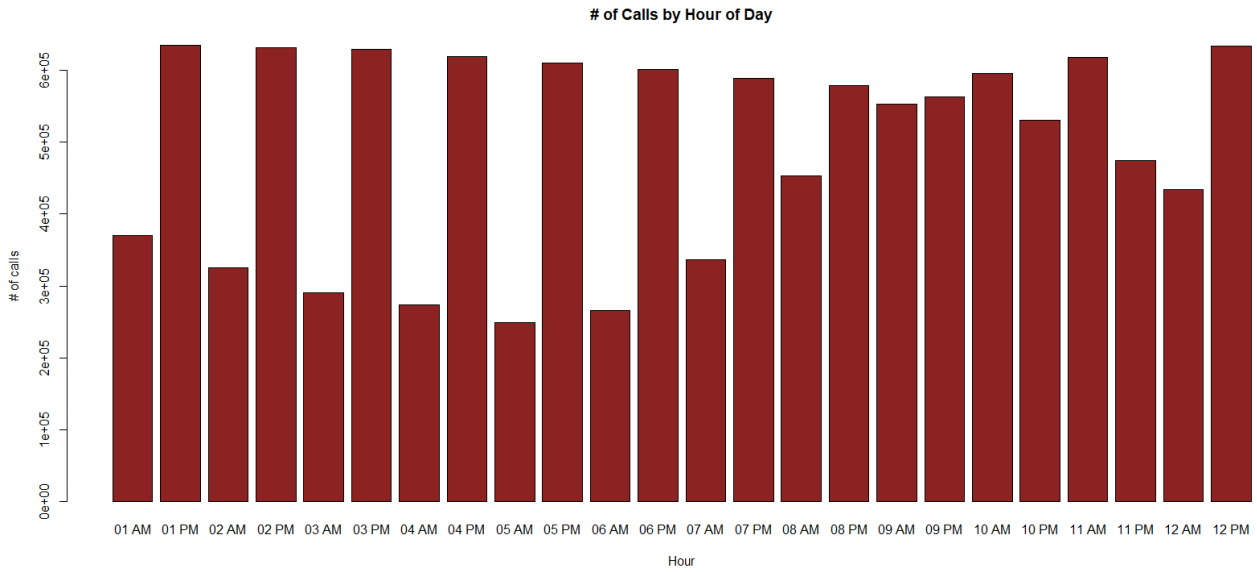


	Month	# of Calls
1	Jan	976155
2	Feb	881258
3	Mar	976679
4	Apr	953244
5	May	1041125
6	Jun	1029786
7	Jul	1056384
8	Aug	1023985
9	Sep	986370
10	Oct	999089
11	Nov	938495
12	Dec	1001189

What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The above bar chart representing the number of calls by Month shows that the Number of calls received in the Month of July are the highest, following by May then June and then December. The graph above also shows that the number of calls in February is lowest when compared to other months and the pattern of number of calls by month is not consistent.

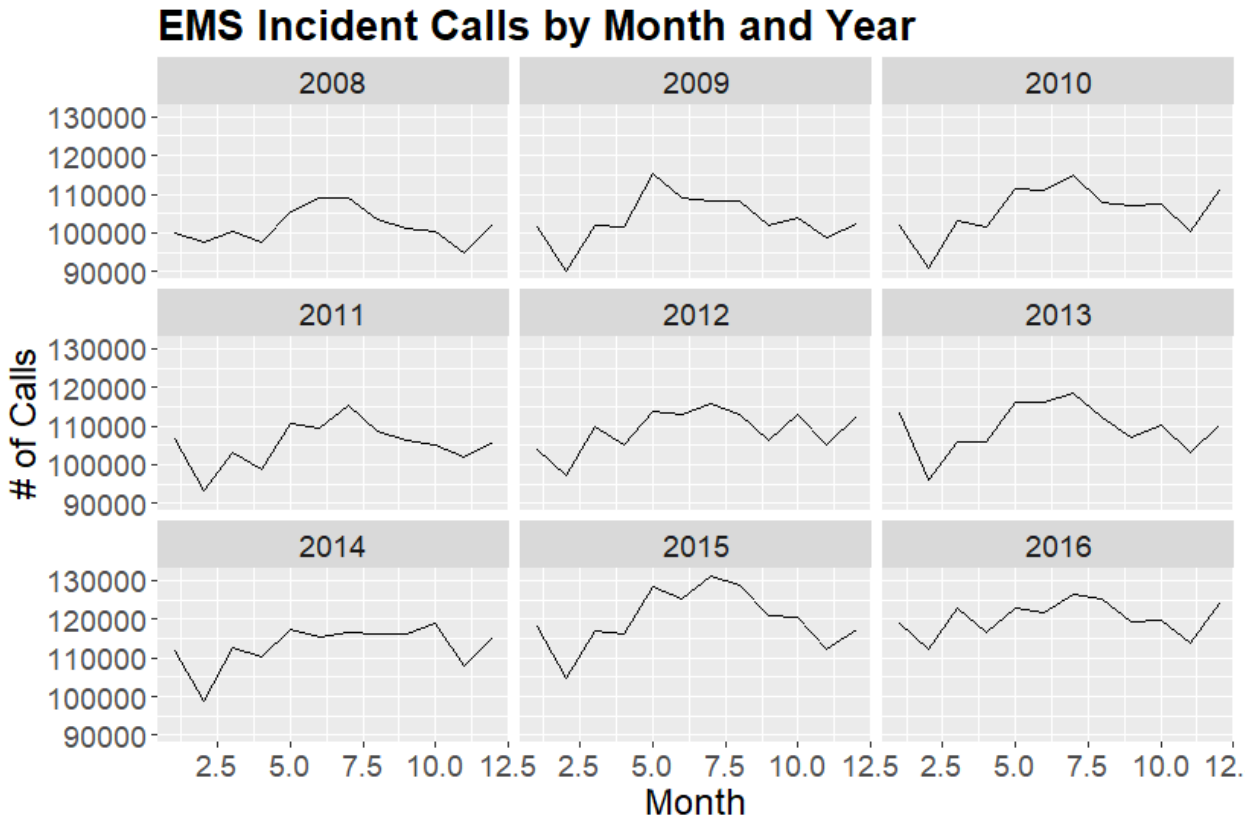
# of calls by hour of day (i.e. 12am-12am)



What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The above bar chart representing the number of calls by Hour of the Day shows that the Number of calls received at 12pm and 1 pm are the highest, followed by 2 pm and 3 pm. The graph above also shows that the number of calls at 5 am is lowest when compared to other hours of the day. The overall graph shows that the number of calls after midnight at 1 pm till the morning at 7 am are comparatively lower than the number of calls received during the daytime from 8 am in the morning till 12 pm in the midnight.

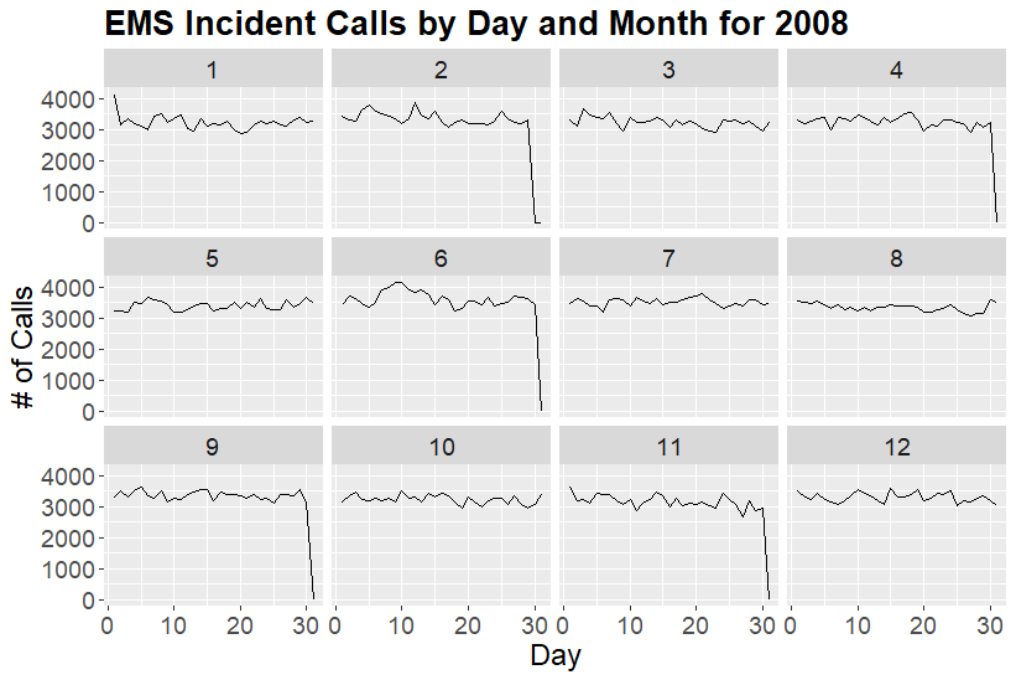
# of calls by month and year (i.e. Jan 2008 – Dec 2016)



What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

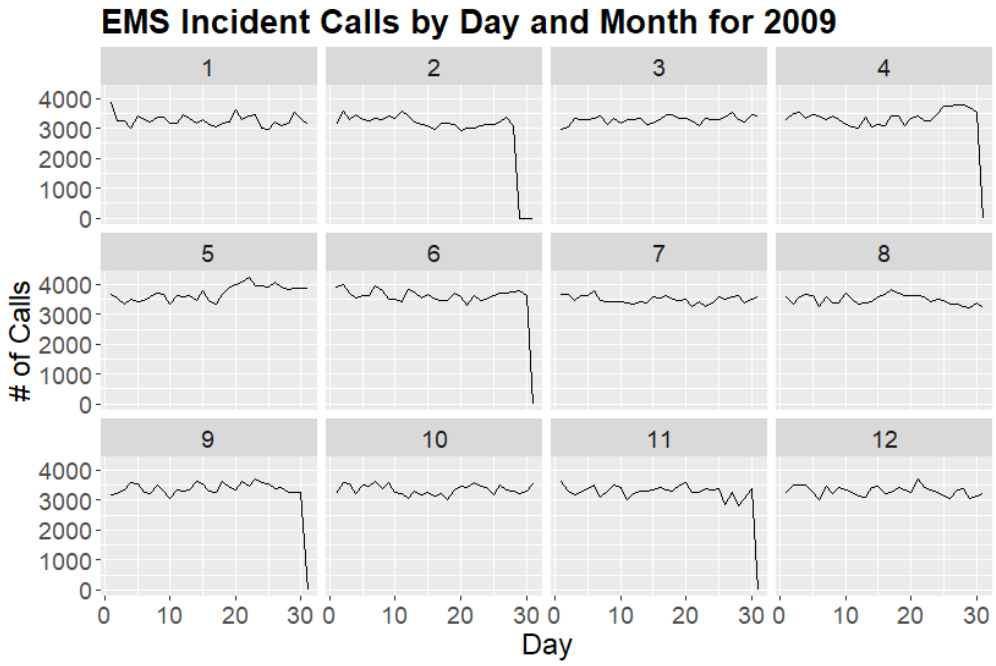
The above graph shows the number of calls by month and year. We can see the highest number of calls in the year 2015 in the months April to July. The year 2008 has the lowest call volume among all the years. The peak call volumes are always around the months of June, July and August, apart from the year 2014, where the peak call volume was around the month October.

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



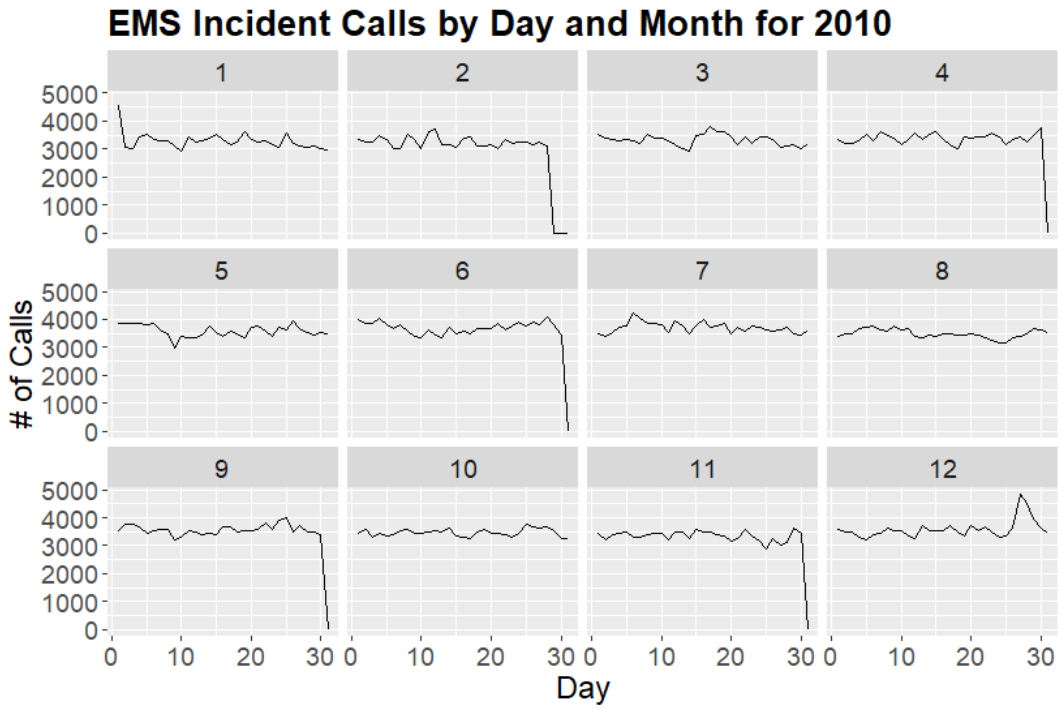
What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The highest peak in the year 2008 is around 8-9 June. January 2008 also started with a very high number of calls.



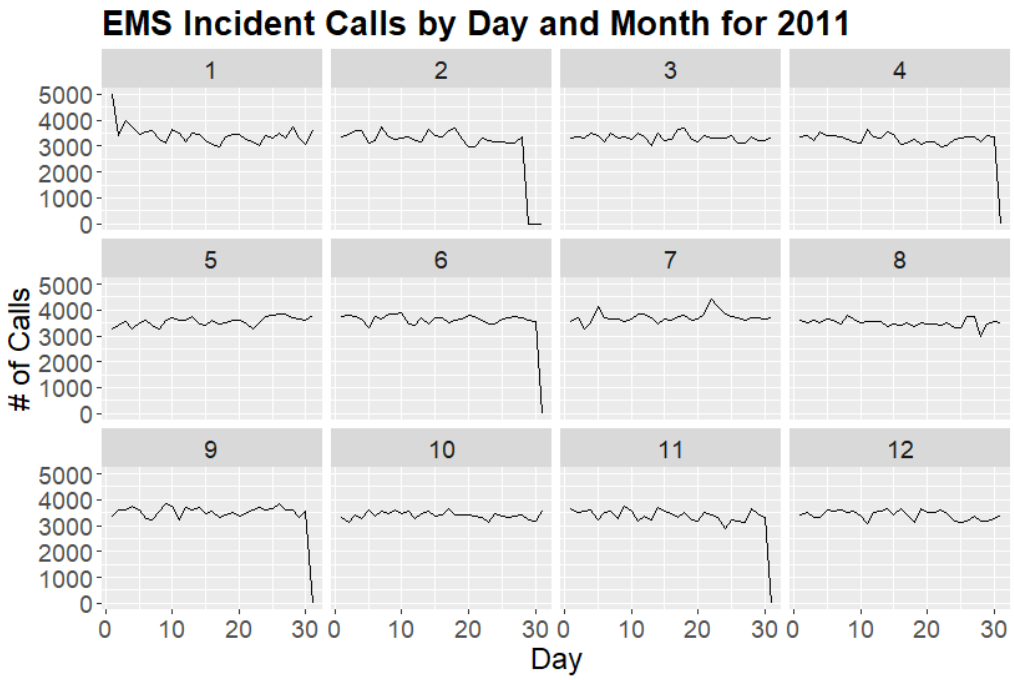
What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The highest peak in the year 2009 is around 23-24 May. In general the months of May and June have more calls than other months in this year.



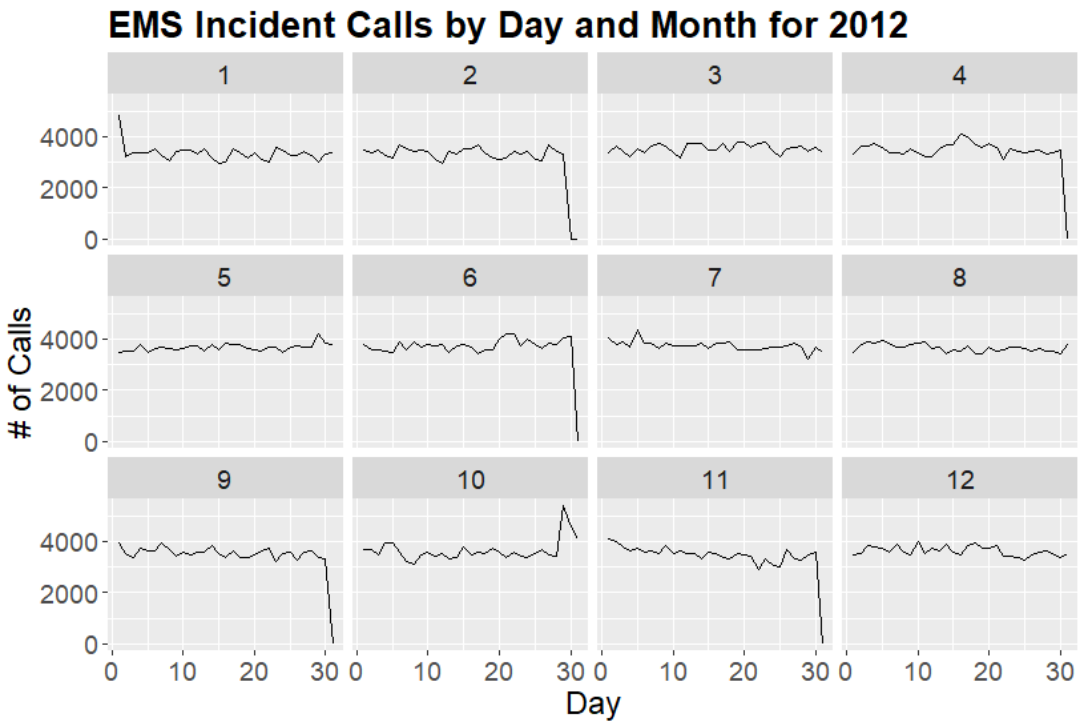
What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

In the Year 2010 January had an abnormally high volume of calls on the 1st day while the number of calls dropped drastically on the subsequent days. We see the highest peak in the next couple of days after Christmas around 26-27 December.



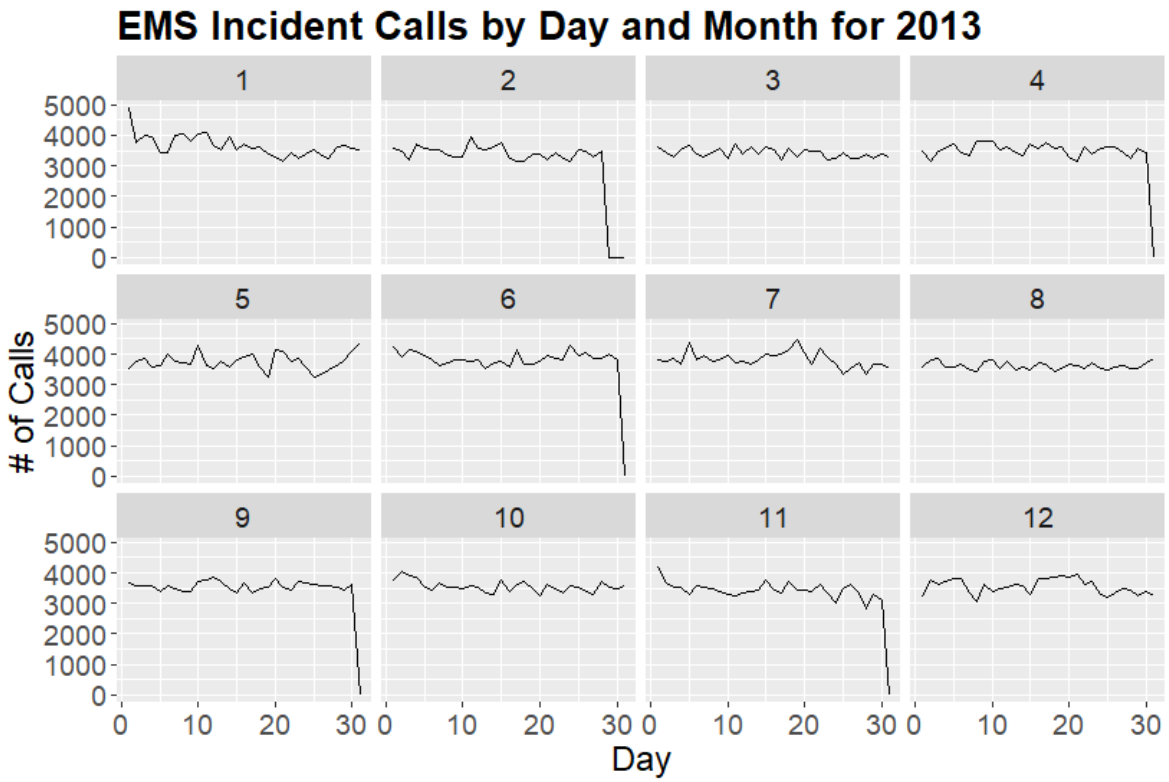
What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The Year 2011 follows the same pattern as 2010 with a very high call volume on 1st January. But the number of calls around 5000 are higher than compared to 2010 which was around 4000. The calls then dropped in the next couple of days. In 2011 the highest call volume was seen on 1st January.



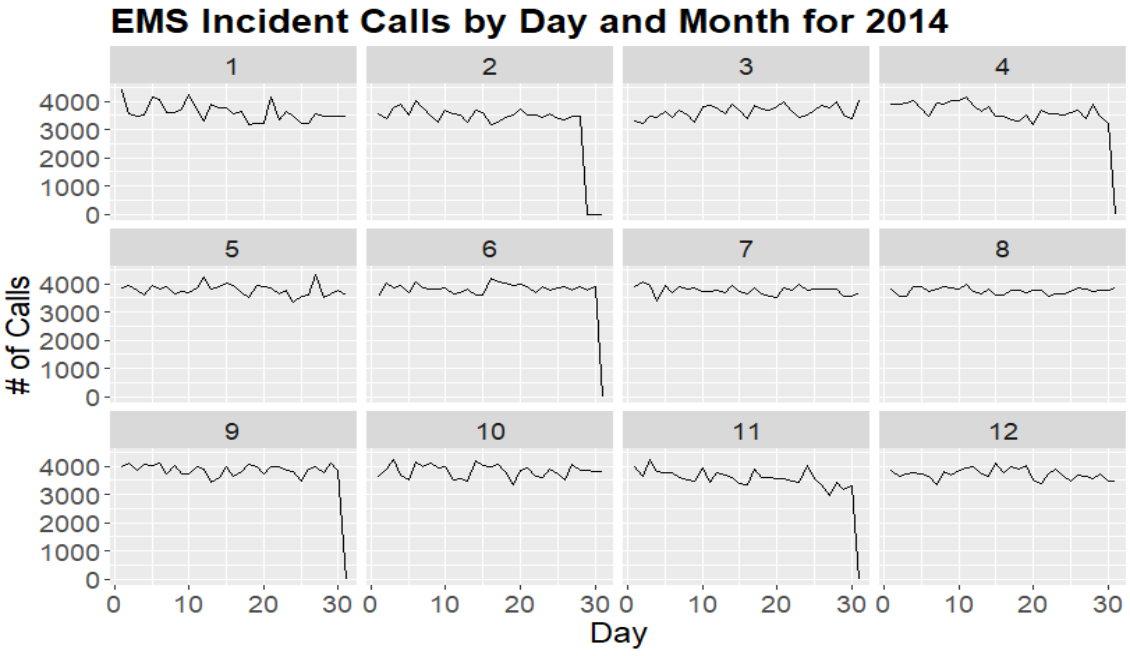
What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

Again in 2012 there was a significant peak on 1st Januray which is a pattern seen in all the years possibly due to New Year's traffic and weather. But there is a significant peak around 29th October in 2012



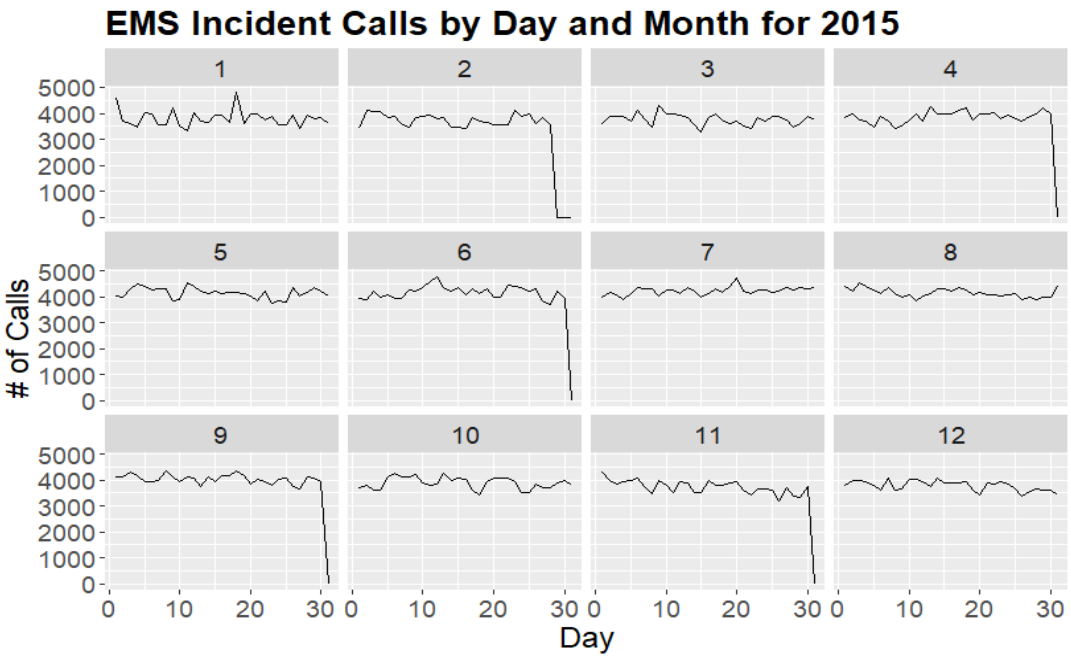
What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

Following the same pattern as last years there is a peak on 1st January 2013 with call volume dropping after.



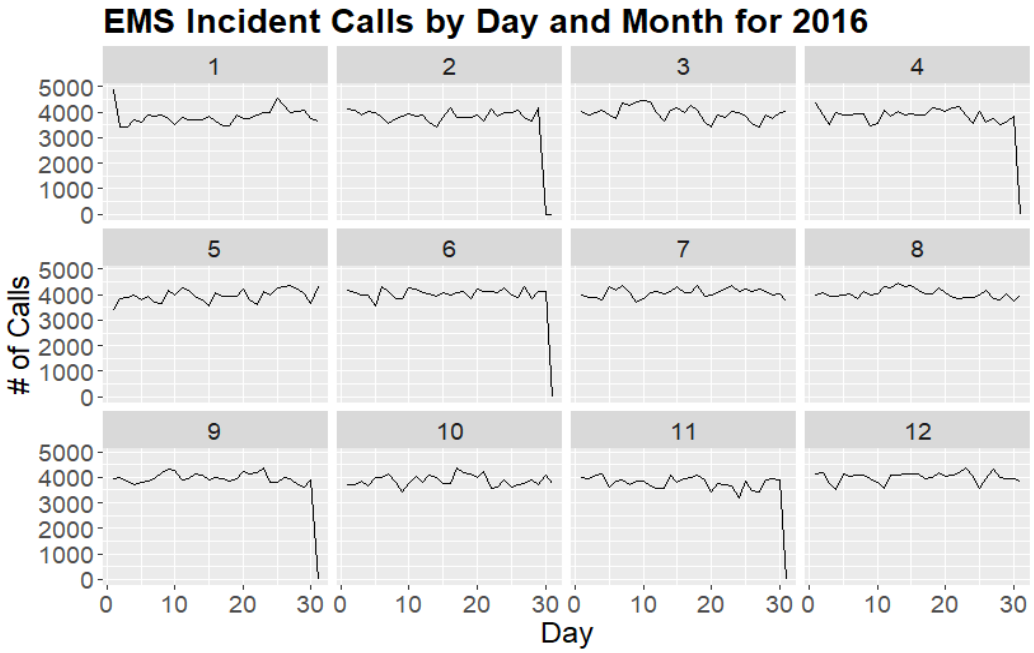
What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

January 2014 had much more call volume throughout the month when compared to other years. 1st January also saw the peak volume of call on any single day in 2014.



What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

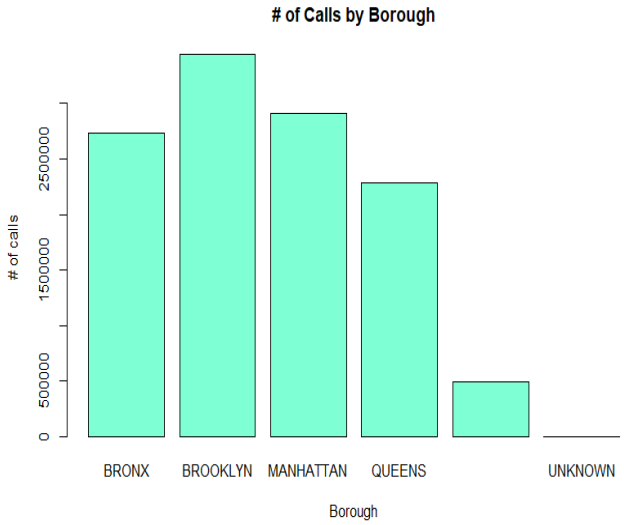
The highest call volume in the year 2015 was around 18th January.



What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

2016 follows the same pattern as every other year with a very high call volume on 1st January and calls dropping after that. The rest of the months of this year show approximately the same pattern.

# of calls by Borough



	Borough	# of Calls
1	BRONX	2733876
2	BROOKLYN	3441280
3	MANHATTAN	2909565
4	QUEENS	2288528
5	RICHMOND / STATEN ISLAND	490386
6	UNKNOWN	124

What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The above graph and table show that the Borough" Brooklyn" receives the greatest number of calls with 3441280, followed by Manhattan with 2909565 number of calls. There is no doubt that these two Boroughs of

New York City are the most popular and crowded, which makes them the more accident-prone areas. Staten Island has the least number of calls with 490386. The unusual thing about the above chart is that there is one unknown borough, which shows the number of calls received as 124, which is the least number compared to all other boroughs. This unknown borough is also acting as an outlier in the above graph.

# of calls by Community District

Community District	No_of_Calls
East New York and Starrett City	378032
Jamaica and Hollis	361438
East Harlem	330581
Midtown	324173
Bedford Stuyvesant	320418
Highbridge and Concourse	312674
Lower East Side and Chinatown	297128
Fordham and University Heights	290914
Parkchester and Soundview	285014
Mott Haven and Melrose	281184
Clinton and Chelsea	268995
Central Harlem	268268
Washington Heights and Inwood	265852
Brownsville	250877
Kingsbridge Heights and Bedford	243956
Fort Greene and Brooklyn Heights	241314
St. George and Stapleton	239086
Morrisania and Crotona	237184
Williamsbridge and Baychester	219268
Upper West Side	217582
Long Island City and Astoria	216216
Belmont and East Tremont	215062
East Flatbush	202899
Greenpoint and Williamsburg	197397
Upper East Side	195272
Queens Village	194163
Coney Island	193533
Stuyvesant Town and Turtle Bay	190582
Flatlands and Canarsie	188851
Flushing and Whitestone	183693
Flatbush and Midwood	182561
Bushwick	181609
Crown Heights and Prospect Heights	175755

Morris Park and Bronxdale	171154
Morningside Heights and Hamilton Heights	168924
Rockaway and Broad Channel	164066
Throgs Neck and Co-op City	154082
Hunts Point and Longwood	149134
Greenwich Village and Soho	148308
Bensonhurst	143670
Jackson Heights	143267
South Crown Heights and Lefferts Gardens	142772
Kew Gardens and Woodhaven	139112
Financial District	135536
South Beach and Willowbrook	134469
Sheepshead Bay	132537
Ridgewood and Maspeth	132217
Elmhurst and Corona	131341
Hillcrest and Fresh Meadows	124772
South Ozone Park and Howard Beach	119779
Park Slope and Carroll Gardens	113724
Sunset Park	113626
Riverdale and Fieldston	112917
Borough Park	112757
Bay Ridge and Dyker Heights	112695
Tottenville and Great Kills	107787
Woodside and Sunnyside	107387
Rego Park and Forest Hills	86781
Bayside and Little Neck	68625

What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

After looking at the above table representing the number of calls by Community District, it shows that the Number of calls received in the East New York and Starrett City district are the highest with 378032. Followed by Jamaica and Hollis with 361438 number of calls. The lowest number of calls 68625 was received by Bayside and Little Neck district. The high number of calls in East New York and Starrett City will affect the incident response time negatively.

# of calls by Initial Call Type

Initial_Call_Type	no_of_Calls
Sick	2049109
Non-Critical Injury	1840886
Difficult Breather	1051494
Psychiatric Patient	843135
Hx Drug Or Alcohol Abuse	684123
Unconscious Patient	657792
Abdominal Pain	560576
Caller Has No Pt Medical Info	460021
Cardiac Condition	452863
Auto Acc W-Injuries	412764
Unknown Condition	264470
Respiratory Distress	220775
Pedestrian Struck	215551
Asthma Attack	195751
Status Epilepticus	183711
Altered Mental Status	161274
Seizures	158732
Cardiac Arrest	147686
Major Injury	123515
Sick Pediatric, <5 Year Old	109519
Internal Bleeding	98303
Minor Injury	94479
CVA (Stroke)	91116
Female In Labor	87489
Anaphylaxis	58733
Stabbing	54241
Multiple Trauma Patient	52761
Stroke	51084
Obstetric Complications	47349
Request For Stand-By	42534
Gun Shot Wound	37397
Miscarriage	31271
Minor Illness	29860
Gyn Bleeding-Pt Not Pregnant	27929
Major Obstetrical Complaint	26642
Hypertension	22773
Reaction To Medication	21772

Choking	20485
Inhalation Of Smoke	17544
Police 10-13, Unconfirmed	15865
One Alarm Fire|	15165
Major Burns 18% Adlt 10% Child	13443
Sick - Cough & Fever	10614
Minor Burns <18% Adlt Or <10%	10376
Rape	10191
Injury Lower Ext In Elderly	9096
Heat Exhaustion	7005
Gyn-Severe Pain-Bleeding	6798
Jumper Up	5698
Auto Accident, No Confirmd Inj	5608
Hypothermia	5594
Jumper Down	5572
Drowning	4802
Fire75 Working Fire	4701
Sick Ped<5 Yrs-Fever & Cough	4084
Child Abuse	3644
Amputation, Fingers Or Toes	3340
Baby Out Or Imminent Birth	2407
Diff Breathing - Fever&Cough	2269
Electrocution	1593
Special Event	1297
Resp Distress - Fever&Cough	968
Abdominal Pain-Fever & Cough|	865
Asthma Attack - Fever&Cough	705
Sick - Rash And Fever	705
One Alarm Fire	598
Stat Transfer Request	472
Venom (Snake Bites)	421
Amputation, Arm, Hand,Leg,Foot	266
Occupied High-Rise Building|	259
Sick Ped<5 Yrs-Rash & Fever	245
Evac	221
Cardiac Condition-Fever&Cough	194
Internal Bleeding-Fever&Cough	117
Report Of Explosives|	89
Medevac, T-C Authority Only	82
ACC	78
Alt Mental Status-Fever&Cough	70

Police 10-13, Confirmed	55
Sick Patient Fever-Travel	55
Seizures - Fever & Cough	52
Difficult Breather Rf	51
Death Confirm By Medical Auth|	47
Two Alarm Fire|	46
Unc Patient - Fever & Cough	36
Fire77 High Rise Residential	28
Stroke - Fever & Cough	27
Abdominal Pain-Fever & Cough	26
Mult Or Prolong Seizur-Fev&Cou	25
Reaction To Med - Fever&Cough	24
Active Shooter	22
Anaphylactic Shock-Fever&Cough	20
Hazardous Materials Incident|	18
BBP	17
Hostage Situation - Barricaded|	15
MOSINJ	14
Fire76 High Rise Commercial	13
Ground Transport Incident|	11
Death Confirm By Medical Auth	10
Stroke Critical - Fever&Cough	9
Structural Collaspe [Specify]	9
Hx Drug Or Alchl Abuse-Fev&Cou	7
ADM	6
Occupied High-Rise Building	6
STUCK	6
Five Alarm Fire Or Greater	5
Report Of Explosives	5
All Other MCIs|	5
Two Alarm Fire	4
Four Alarm Fire	4
Aircraft Incident - Crash	4
Choking Fever&Cough	3
Three Alarm Fire	3
Rapid Transit-Rail Incident	3
Power Failure - Blackout	3
MOSILL	3
Unconscious Fever-Travel Patient	3
Unconscious Patient-Rash&Fever	3
Card Or Resp Arrest-Fevercough	2

Difficult Breathing Fever-Travel	2
MCI25	2
Construction-Demolition Incid|	2
MECHE	2
Respiratory Distress Fever-Travel	2
Test Kdt-Modat	2
Abdominal Pain Fever-Travel	1
Asthma Patient Fever-Travel	1
DRILL	1
Criminal Detection Facil Incid	1
Explosion	1
Ground Transport Incident	1
Confined Space Incident	1
MCI37	1
Civil Distrubance	1
Hostage Situation - Barricaded	1
Active Shooter|	1
Active Shooter||	1
MECHV	1
NOVEH	1
RADIO	1
Status Epilepticus Fever-Travel	1

What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The above table shows that the initial call type “Sick” has the highest value of 2049109, which means many people call because of sickness. Followed by “Non-Critical Injury "with 1840886 number of calls, then “Difficult Breather”, with 1051494 number of calls and then “Psychiatric Patient” with 843135 number of calls. There are some initial call types that EMS receives the lowest calls for like Seizures - Fever & Cough, Active Shooter Power Failure – Blackout, MCI25, Abdominal Pain Fever-Travel etc.

# of calls by Final Call Type

counts_final_calltype	**no_of_Calls**
Sick	1925873
Non-Critical Injury	1799152
Difficult Breather	1100885
Psychiatric Patient	808263
Hx Drug Or Alcohol Abuse	713540

Unconscious Patient	689235
Abdominal Pain	557298
Cardiac Condition	525473
Auto Acc W-Injuries	399053
Caller Has No Pt Medical Info	390175
Unknown Condition	221032
Altered Mental Status	217580
Pedestrian Struck	215638
Cardiac Arrest	213173
Respiratory Distress	207017
Status Epilepticus	202481
Asthma Attack	187808
Seizures	144941
Major Injury	138544
Sick Pediatric, <5 Year Old	108968
Internal Bleeding	105697
Minor Injury	92491
CVA (Stroke)	87309
Female In Labor	84578
Anaphylaxis	62877
Multiple Trauma Patient	59788
Stabbing	56913
Obstetric Complications	50497
Stroke	47994
Request For Stand-By	38488
Gun Shot Wound	37156
Miscarriage	31362
Minor Illness	29226
Major Obstetrical Complaint	28582
Gyn Bleeding-Pt Not Pregnant	28043
Choking	22145
Hypertension	21685
Reaction To Medication	21140
One Alarm Fire|	20318
Inhalation Of Smoke	18137
Major Burns 18% Adlt 10% Child	14933
Police 10-13, Unconfirmed	14479
Sick - Cough & Fever	10830
Rape	10584
Injury Lower Ext In Elderly	10373
Minor Burns <18% Adlt Or <10%	10044

Gyn-Severe Pain-Bleeding	7505
Heat Exhaustion	6975
Jumper Up	6462
Jumper Down	6183
Hypothermia	5951
Auto Accident, No Confirmd Inj	5189
Drowning	4883
Sick Ped<5 Yrs-Fever & Cough	4439
Baby Out Or Imminent Birth	4021
Child Abuse	3882
Amputation, Fingers Or Toes	3771
Diff Breathing - Fever&Cough	3049
Ground Transport Incident|	2563
Hostage Situation - Barricaded|	2060
Electrocution	1725
Special Event	1324
Two Alarm Fire|	1316
Resp Distress - Fever&Cough	1173
Abdominal Pain-Fever & Cough|	1023
Asthma Attack - Fever&Cough	806
Report Of Explosives|	769
Sick - Rash And Fever	746
Hazardous Materials Incident|	708
Occupied High-Rise Building|	691
Stat Transfer Request	504
Venom (Snake Bites)	456
Amputation, Arm, Hand,Leg,Foot	371
Sick Ped<5 Yrs-Rash & Fever	276
Three Alarm Fire|	261
One Alarm Fire	260
Police 10-13, Confirmed	243
Cardiac Condition-Fever&Cough	230
Evac	224
All Other MCIs|	208
Internal Bleeding-Fever&Cough	151
Structural Collaspe [Specify]|	137
Alt Mental Status-Fever&Cough	113
Confined Space Incident	112
Medevac, T-C Authority Only	92
Four Alarm Fire|	84
Construction-Demolition Incid|	84

Unc Patient - Fever & Cough	83
Death Confirm By Medical Auth	72
Difficult Breather Rf	69
Seizures - Fever & Cough	67
Five Alarm Fire Or Greater	58
Mult Or Prolong Seizur-Fev&Cou	51
Marine - Harbor Incident|	48
Stroke - Fever & Cough	44
Fire75 Working Fire	39
Sick Patient Fever-Travel	39
Rapid Transit-Rail Incident|	29
Reaction To Med - Fever&Cough	29
Civil Distrubance	28
Abdominal Pain-Fever & Cough	27
Explosion	26
Anaphylactic Shock-Fever&Cough	24
Criminal Detection Facil Incid|	24
Aircraft Incident - Crash	20
Active Shooter	18
Stroke Critical - Fever&Cough	16
Card Or Resp Arrest-Fevercough	14
Hostage Situation - Barricaded	14
Hazardous Materials Incident	14
Hx Drug Or Alchl Abuse-Fev&Cou	13
Medical Facility Evacuation	11
Power Failure - Blackout|	10
Ground Transport Incident	9
All Other MCIs	6
Report Of Explosives	5
Unconscious Fever-Travel Patient	5
Choking Fever&Cough	4
Abdominal Pain Fever-Travel	3
Difficult Breathing Fever-Travel	3
Two Alarm Fire	3
Occupied High-Rise Building	3
Active Shooter||	2
Unconscious Patient-Rash&Fever	2
ALMNFC	1
ARSTFC	1
Fire77 High Rise Residential	1
MCI27	1

Structural Collaspe [Specify]	1
MCI37	1
MCI42	1
Respiratory Distress Fever-Travel	1
Status Epilepticus Fever-Travel	1

What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

After looking at the above table representing the number of calls by Final Call Type, it shows that it follows the same pattern as the number of calls by initial call type with highest number of calls for sickness, Non-Critical Injury, Difficult Breather, Psychiatric Patient etc. However, there are some differences in the number of calls for each type.

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

initial_severity_level_code	1	2	3	4	5	6	7	8
0	0	0	0	0	0	1	0	0
1	164262	1418	735	1131	317	734	153	7
2	49054	2278633	8850	6458	2174	7619	1123	259
3	8628	54845	1424569	6032	5630	8294	2902	104
4	6200	71274	46332	2317895	15232	12876	3308	189
5	2587	27066	52970	37179	2141574	2152	3696	99
6	3870	42888	81912	44584	27471	1833836	8510	682
7	745	6515	9306	27108	11514	6709	938364	83
8	48	3819	371	454	767	232	211	39025
9	10	25	15	59	24	12	3	26

What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

After looking at the above table representing number of calls, by Initial Severity Level and Final Severity Level, it shows that the severity code from 1 to 7 has the highest number of calls. Whereas the severity code 8 and 9 has significantly a smaller number of calls, which indicates the there are a smaller number of calls with high severity code.

Most of the calls of 9 Initial severity are reduced to 5 or less as depicted in the last row of the table. However we see that most of the calls with 8 Initial Severity level are also assigned 8 Final severity.

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.

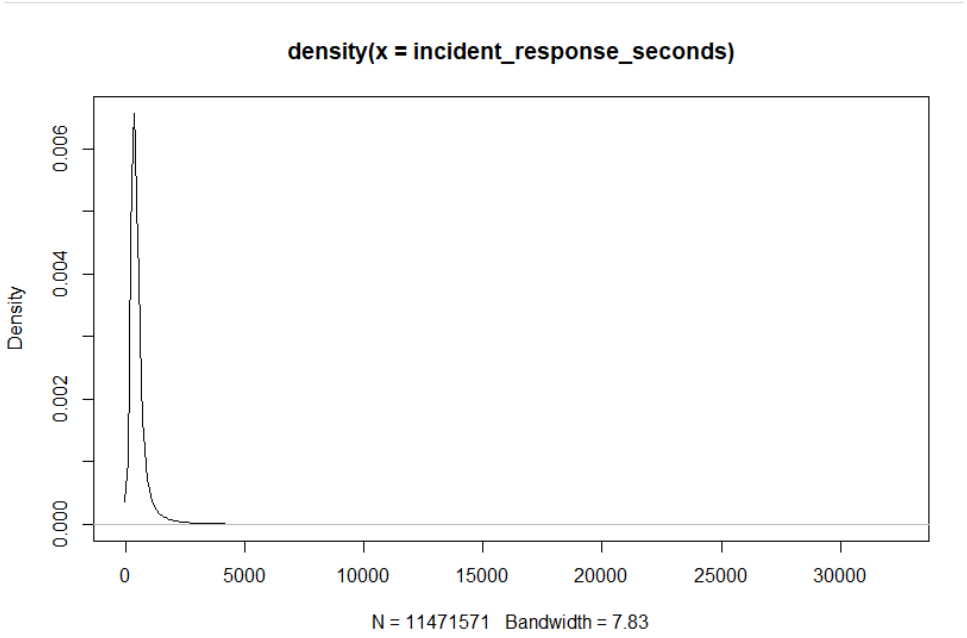
Please complete all sections, and ensure that the analyses provided are formatted, well-described, and clear.

Part 2b – Response Time Analysis

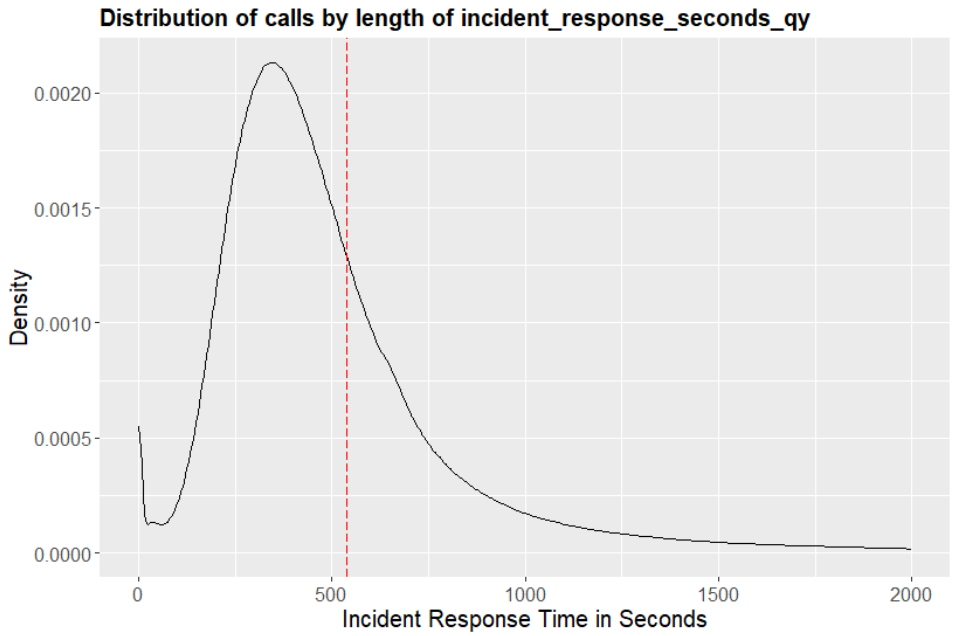
For those who are doing the response time analysis, please complete the following descriptive statistic tables and/or charts. After each group of charts or tables, I want you to include the following section:

Here are the charts to complete:

1. **Distribution of calls by length of incident_response_seconds_qy**



Upon inspection of the density plot of the incident_response_seconds_qy column, we can see that most of the values lie between 0 to 2000 seconds. So, to get a clearer picture, we can remove all rows where incident_response_seconds_qy > 2000.

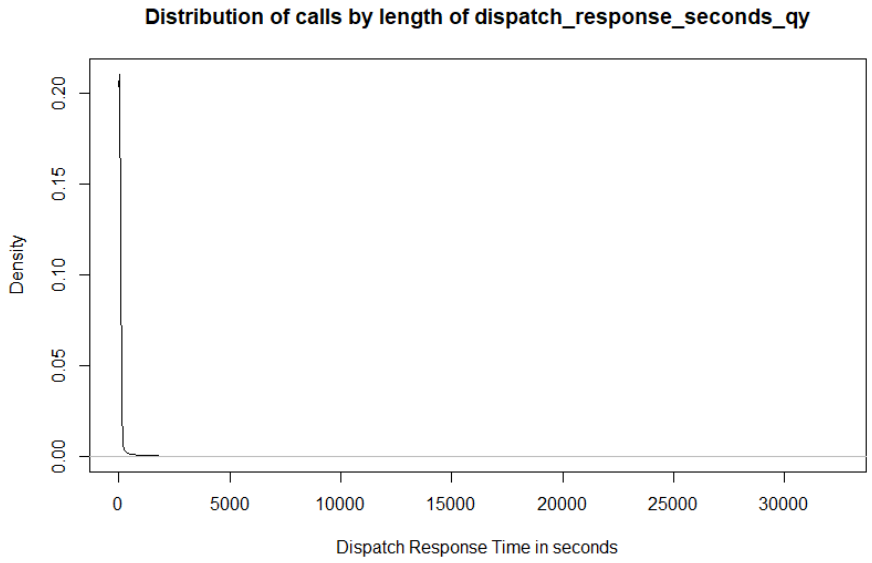


What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

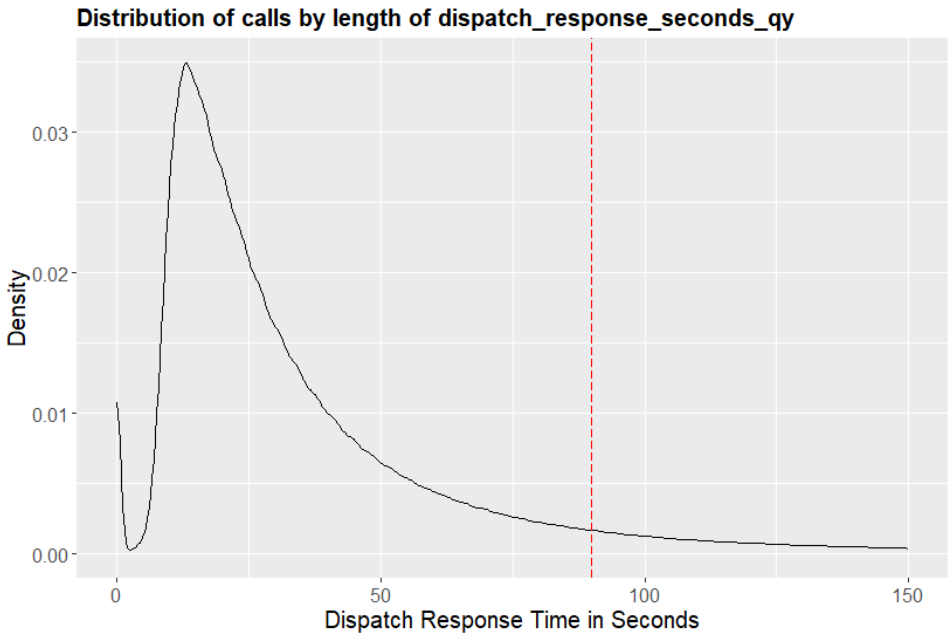
The above density curve shows the Distribution of calls by length of incident_response_seconds_qy. We can see that the curve is highly left skewed with the mean value of approximately 536, which means the mean is less than the median for this curve.

Also, the curve is unimodal, which means it only has one peak, and the highest peak is shown at approximately 375 seconds. We can say that most of the incident response time values lie in the range of approximately 250 to 500 seconds.

2. **Distribution of calls by length of dispatch_response_seconds_qy**



Upon inspection of the density plot of the dispatch_response_seconds_qy column, we can see that most of the values lie between 0 to less than 150 seconds. So in order to get a clearer picture, we can remove all rows where dispatch_response_seconds_qy > 150.

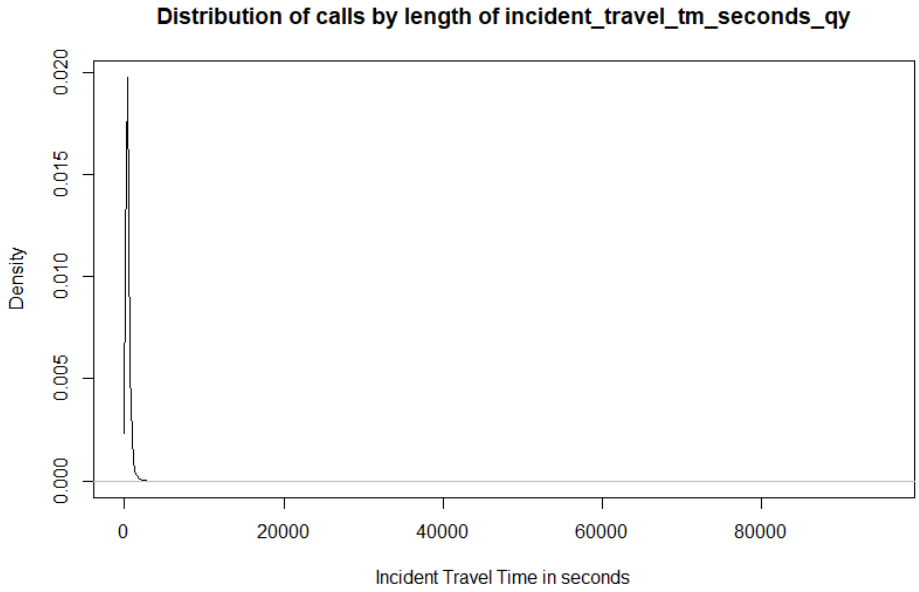


What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

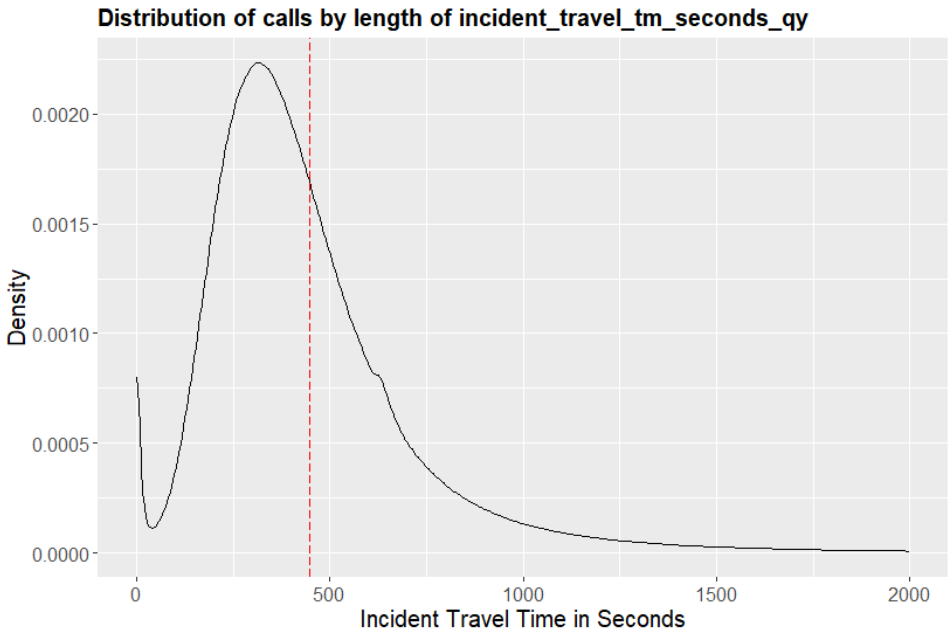
The above density curve shows the Distribution of calls by length of dispatch_response_seconds_qy. We can see that the curve is highly left skewed with the mean value of approximately 90, which means the mean is less than the median for this curve.

Also, the curve is unimodal, which means it only has one peak, and the highest peak is shown at approximately 13 seconds. We can say that most of the dispatch response time values lie in the range of approximately 5 to 50 seconds.

3. **Distribution of calls by length of incident_travel_tm_seconds_qy**



Upon inspection of the density plot of the incident_travel_tm_seconds_qy column, we can see that most of the values lie between 0 to less than 2000 seconds. So, to get a clearer picture, we can remove all rows where incident_travel_tm_seconds_qy > 2000.

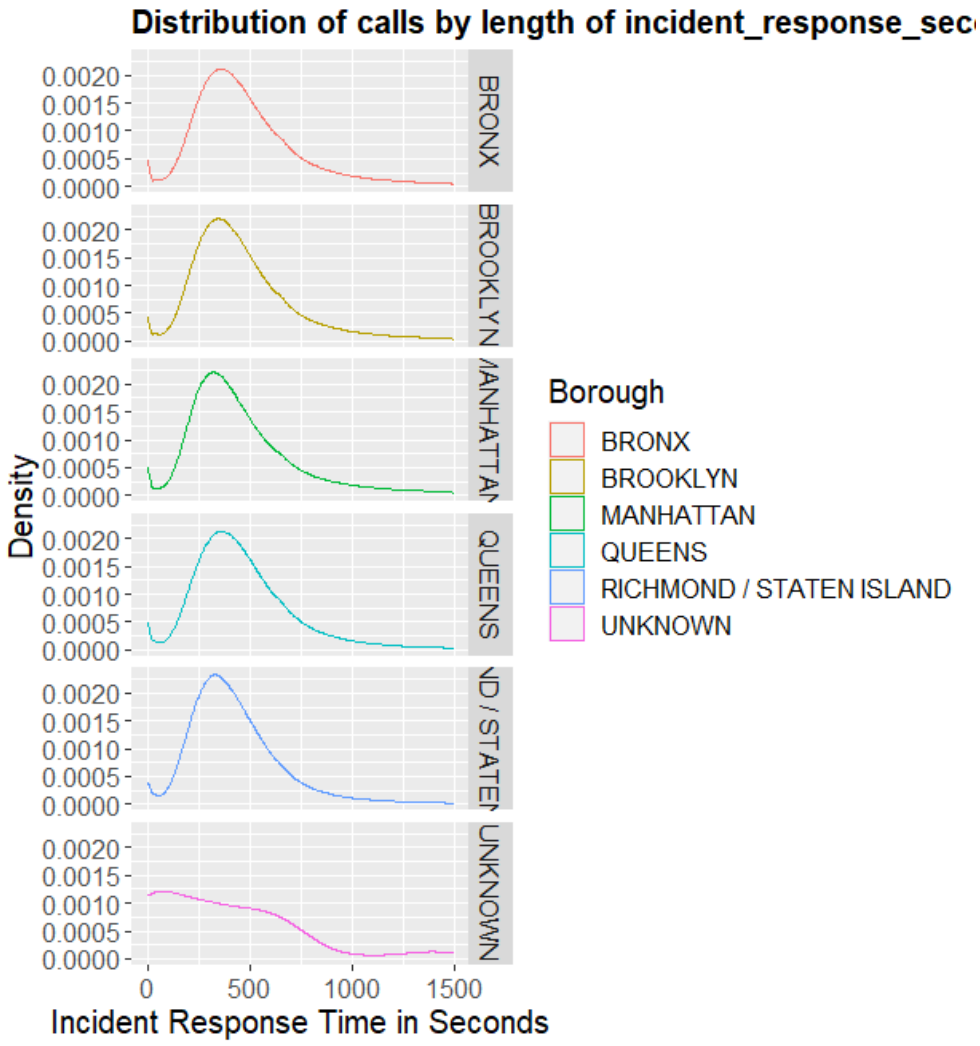


What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The above density curve shows the Distribution of calls by length of incident_travel_tm_seconds_qy. We can see that the curve is highly left skewed with the mean value of approximately 448, which means the mean is less than the median for this curve.

Also, the curve is unimodal, which means it only has one peak, and the highest peak is shown at approximately 325 seconds. We can say that most of the dispatch response time values lie in the range of approximately 100 to 750 seconds.

4. **Distribution of calls by length of incident_response_seconds_qy, for each Borough (i.e. five separate distributions)**

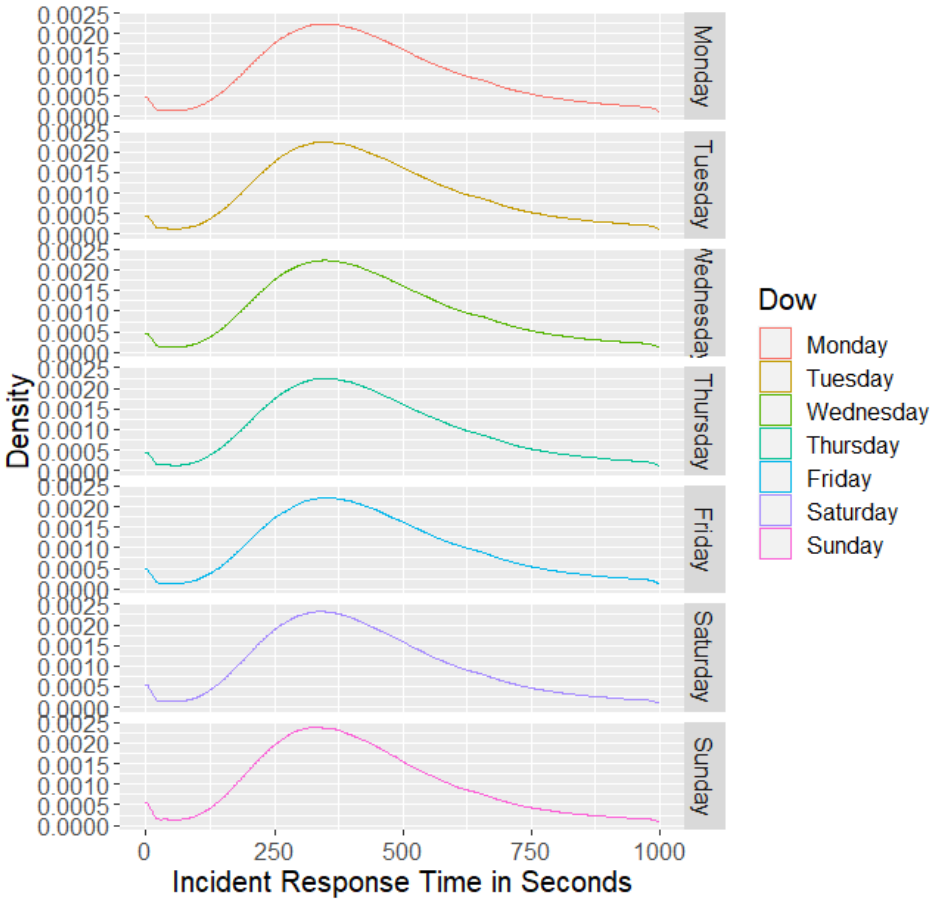


	Borough	Mean Incident Response Seconds
1	BRONX	481.8199
2	BROOKLYN	465.5724
3	MANHATTAN	462.7556
4	QUEENS	467.9174
5	RICHMOND / STATEN ISLAND	429.6148
6	UNKNOWN	361.0396

What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The above Distribution of calls by length of incident_response_seconds_qy, for each Borough shows that almost all the boroughs have approximately same number calls by length of incident response time in seconds. However, we can notice a slight difference in the distribution curve of Richmond/Saten Island borough as it is slightly more left skewed than the rest of the boroughs.

5. **Distribution of calls by length of incident_response_seconds_qy, for day of the week (i.e. seven separate distributions)**

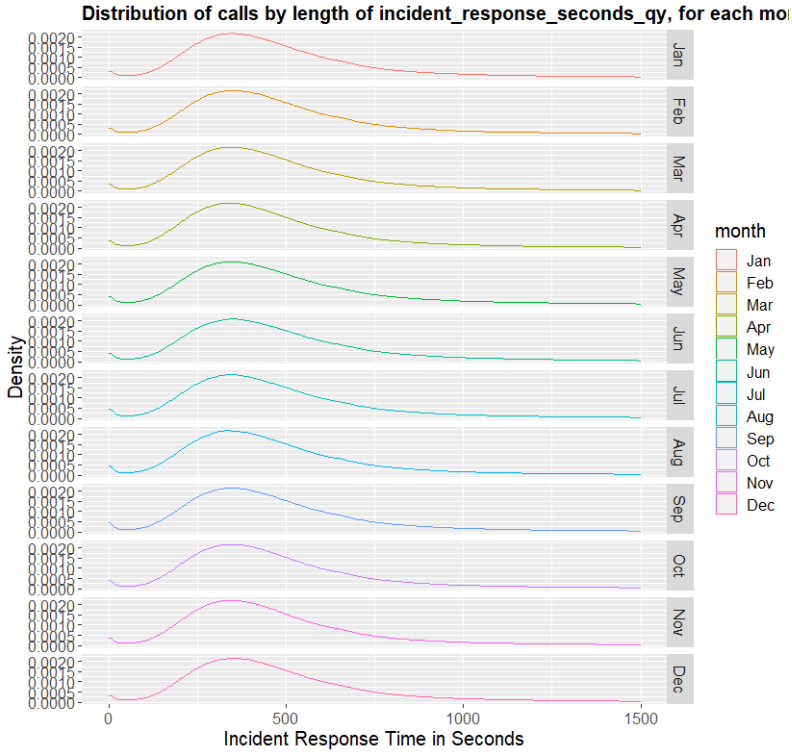


	Dow	Mean Incident Response Seconds
1	Monday	437.1232
2	Tuesday	437.0496
3	Wednesday	436.5334
4	Thursday	437.4596
5	Friday	439.7468
6	Saturday	422.4588
7	Sunday	415.7949

What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The above Distribution of calls by length of incident_response_seconds_qy, for day of the week, shows that almost all the days of week have approximately same number calls by length of incident response time in seconds. However, we can notice a slight difference in the distribution curve of Sunday as it is slightly more left skewed than the rest of the days of the week.

6. **Distribution of calls by length of incident_response_seconds_qy, for each month (i.e. 12 separate distributions)**

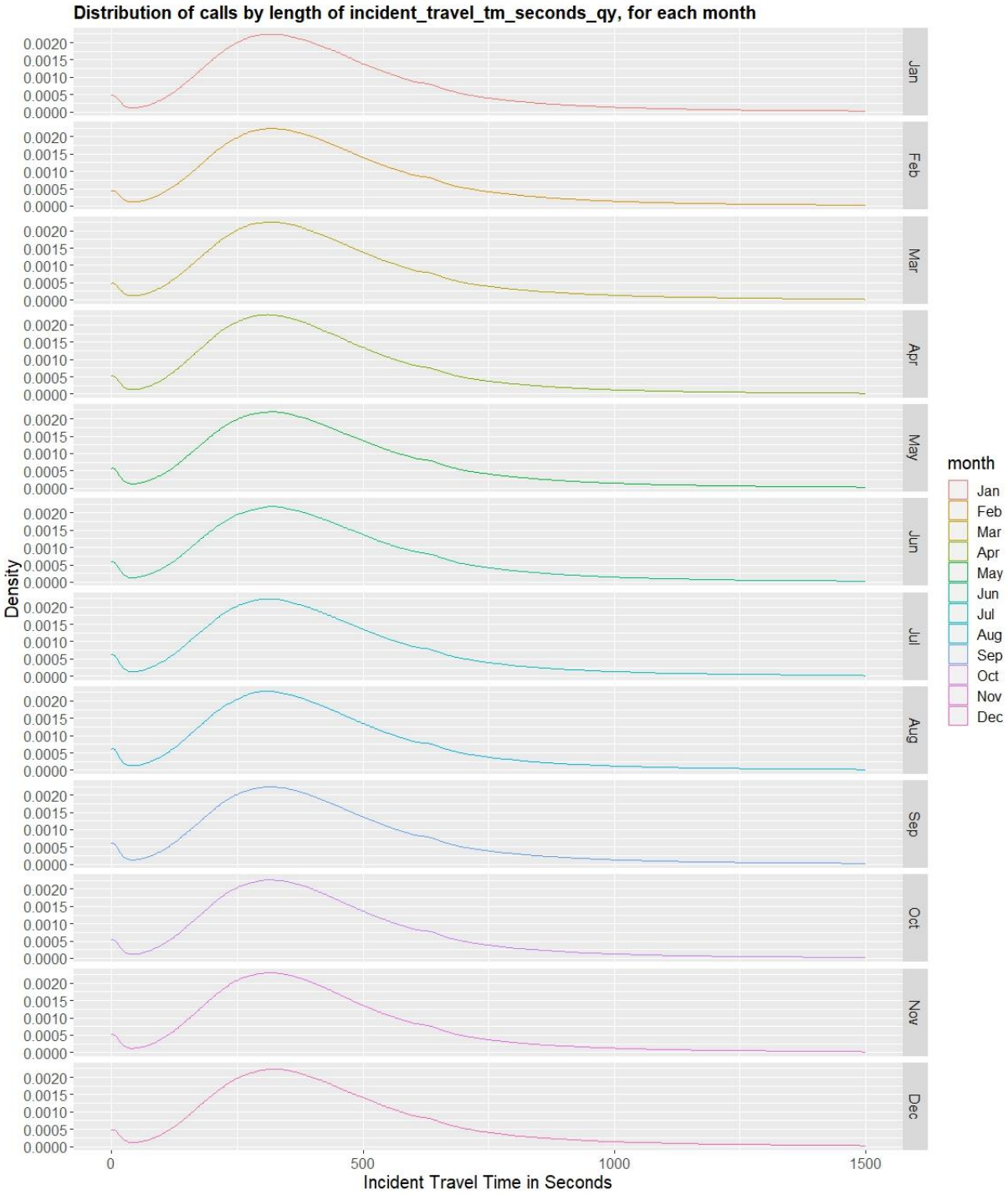


	month	Mean Incident Response Seconds
1	Jan	469.7680
2	Feb	471.5642
3	Mar	468.1668
4	Apr	459.1398
5	May	473.3602
6	Jun	478.3016
7	Jul	467.7155
8	Aug	461.1287
9	Sep	466.3037
10	Oct	464.1684
11	Nov	458.5911
12	Dec	471.9935

What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The above Distribution of calls by length of incident_response_seconds_qy, for each month, shows that almost all the months have approximately the same distribution of calls by length of incident response time in seconds.

7. **Distribution of calls by length of incident_travel_tm_seconds_qy, for each month (i.e. 12 separate distributions)**


Distribution of calls by length of incident_travel_tm_seconds_qy, for each month
Density
0.0020
0.0015
0.0010
0.0005
0.0000
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
0
500
1000
1500
Incident Travel Time in Seconds
month
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

	month	Mean Incident Travel Time Seconds
1	Jan	429.9895
2	Feb	432.7762
3	Mar	427.8050
4	Apr	418.9746
5	May	430.5361
6	Jun	433.4165
7	Jul	423.4674
8	Aug	418.5401
9	Sep	424.6885
10	Oct	423.4884
11	Nov	419.8031
12	Dec	433.3060

What patterns are you seeing? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The above Distribution of calls by length of incident_travel_tm_seconds_qy, for each month, shows that almost all the months have approximately the same distribution of calls by length of incident travel time in seconds. Looking at the mean June and December have slightly higher Incident Travel Time in seconds than other months.

Part 3 – Additional Data Analysis

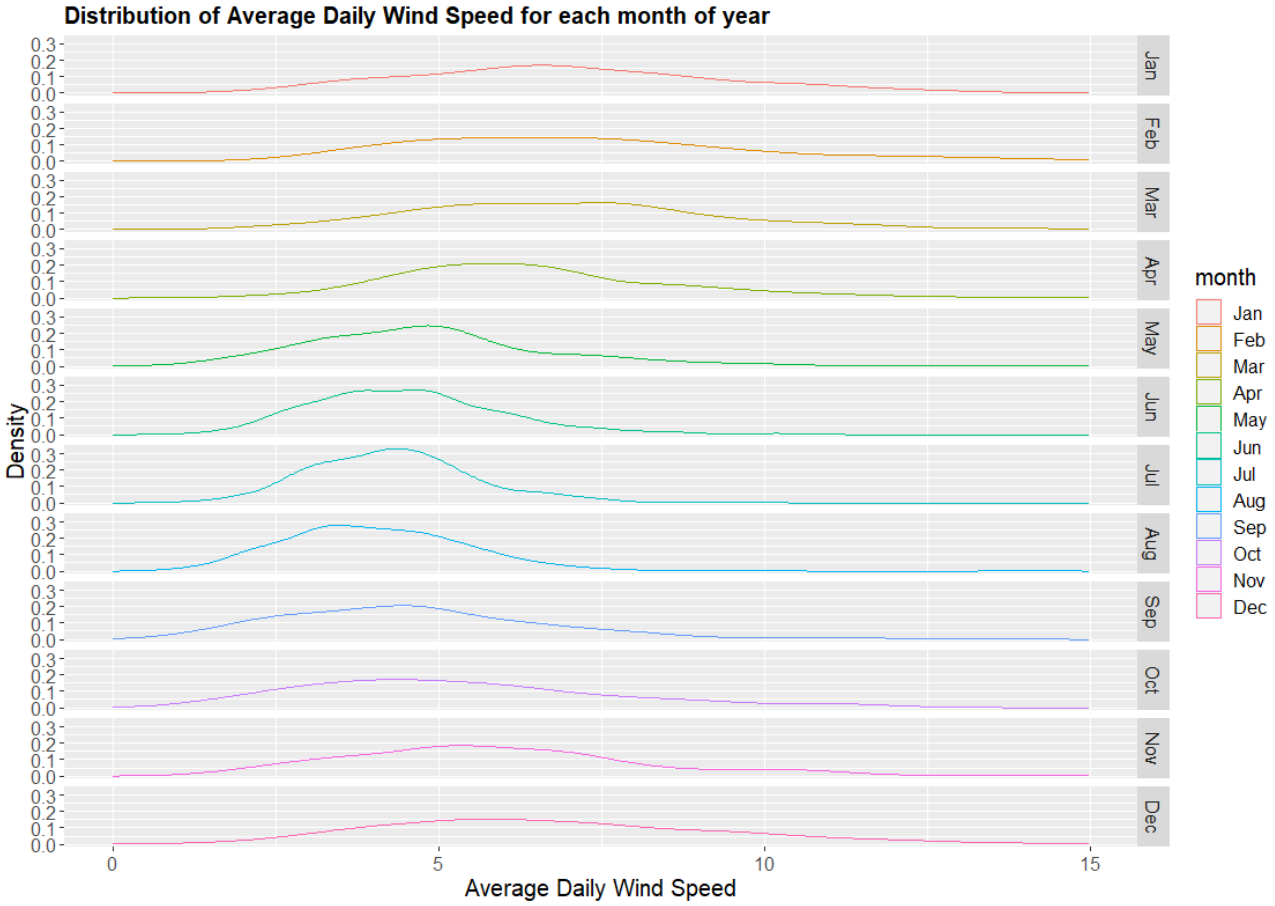
Now that you have completed your analysis of the overall data and your specific task requirements, it is time to focus on the additional data you are planning on using for the modeling. This could be weather data, this could be community district data, or other data sources.

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).

Distribution of Average Daily Wind Speed for each month



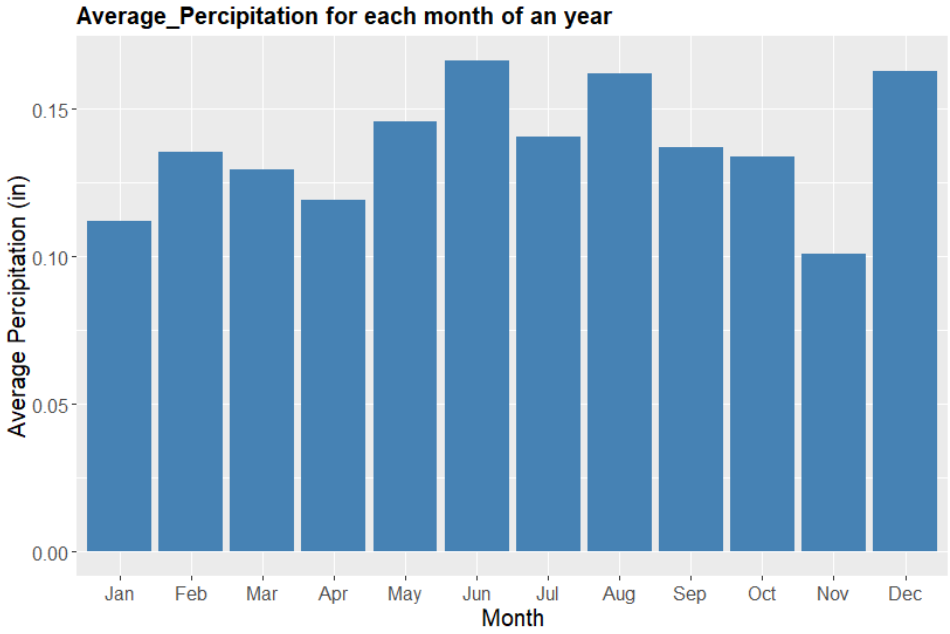
	Month	mean_AWND
1	Jan	6.976089
2	Feb	7.266957
3	Mar	7.056559
4	Apr	6.455556
5	May	4.944440
6	Jun	4.534794
7	Jul	4.321434
8	Aug	4.074493
9	Sep	4.704179
10	Oct	5.348417
11	Nov	5.894142
12	Dec	6.781449

What patterns are you seeing in this additional data you are including in the model? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The above density curve shows the Distribution of Average Daily Wind Speed for each month of the year. We can see that almost all the curves for all the months are left skewed. Especially, the months, May to August, are highly left skewed, which means the mean is less than the median for these curves.

March is somewhat skewed right, but not a significant right. Almost all the curves are unimodal, which means they only have one peak, and the highest peak is shown by the month of July.

Precipitation



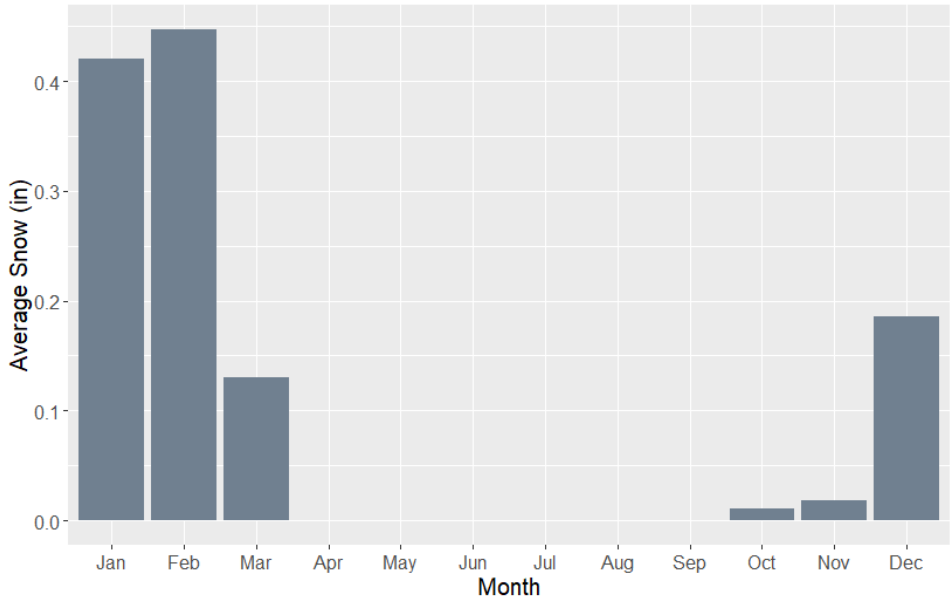
	Month	Average_Percipitation
1	Jan	0.1120430
2	Feb	0.1356078
3	Mar	0.1294982
4	Apr	0.1192963
5	May	0.1457348
6	Jun	0.1663704
7	Jul	0.1405735
8	Aug	0.1620789
9	Sep	0.1370000
10	Oct	0.1337276
11	Nov	0.1008889
12	Dec	0.1628315

What patterns are you seeing in this additional data you are including in the model? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The above graph shows the average precipitation for each month of the year. June, on average for each year, has the highest precipitation with 0.1663704. Followed by December and August with 0.1628315 and 0.1620789. There is an unusual drop in the average precipitation in November, which is the least precipitated month among all.

Snowfall (in)

Average Snowfall for each month

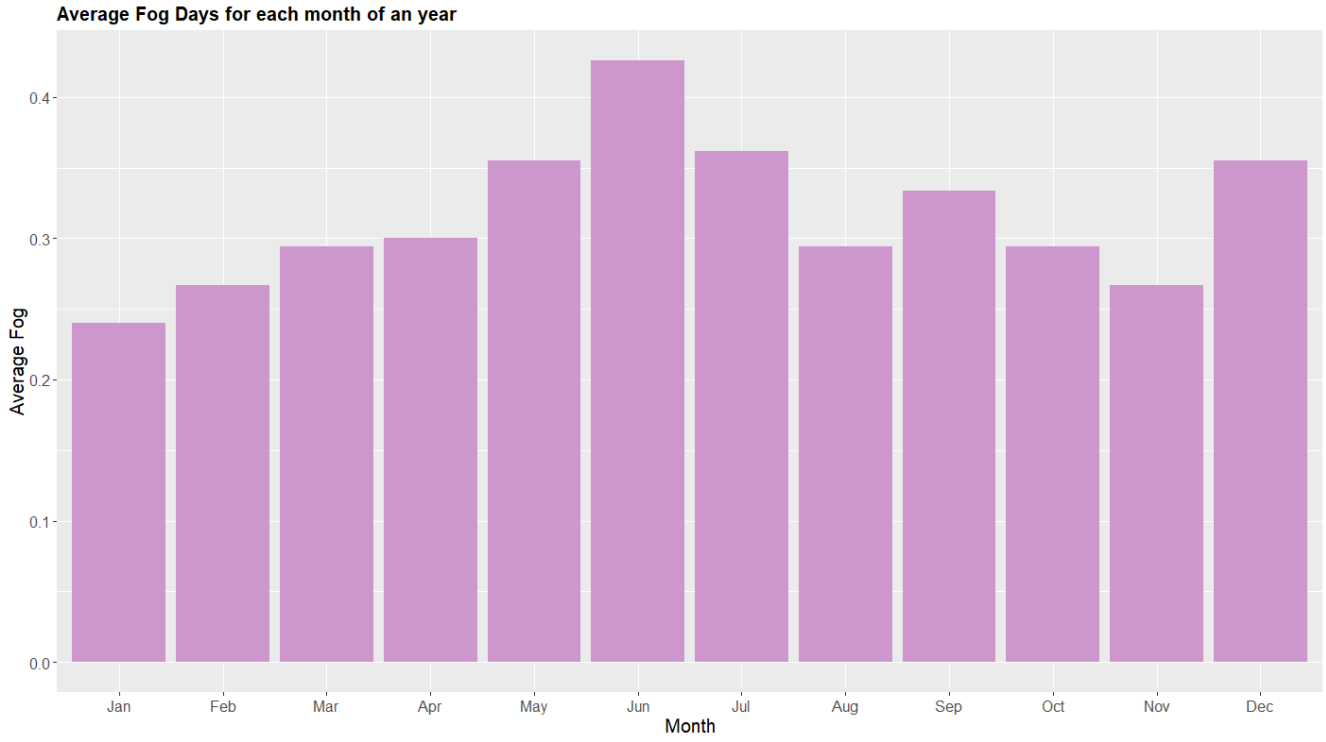


	Month	SNOW_Percipitation
1	Jan	0.42078853
2	Feb	0.44705882
3	Mar	0.12974910
4	Apr	0.00000000
5	May	0.00000000
6	Jun	0.00000000
7	Jul	0.00000000
8	Aug	0.00000000
9	Sep	0.00000000
10	Oct	0.01039427
11	Nov	0.01814815
12	Dec	0.18530466

What patterns are you seeing in this additional data you are including in the model? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The above graph shows the average snowfall for each month of the year. As we know, winter slowly starts in October. We can see a little snow in October and November, which shows the start of snowfall. The first heavy snowfall began in December, which was 0.18530466 inches. We can also see the two February and January, the two months with the highest average snowfall with 0.44705882 and 0.42078853 respectively. In March the average snowfall starts to drop, that indicates the end of snow season.

Fog



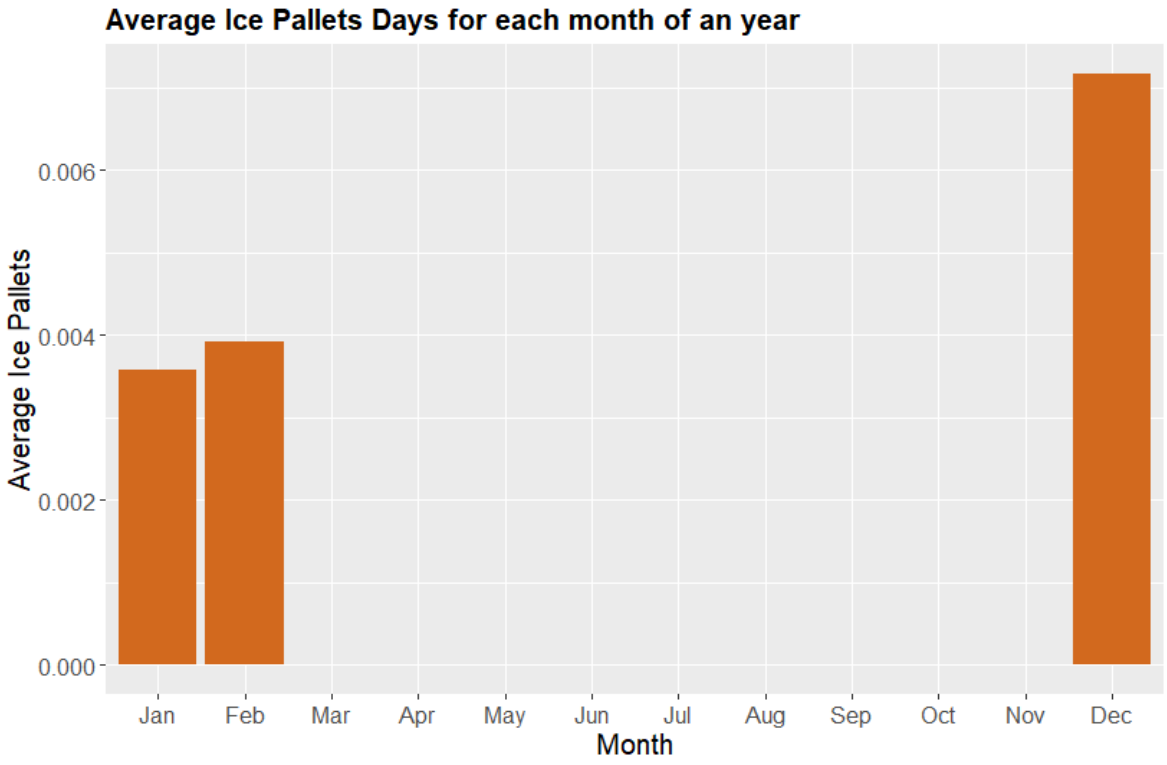
	Month	Fog
1	Jan	0.2401434
2	Feb	0.2666667
3	Mar	0.2939068
4	Apr	0.3000000
5	May	0.3548387
6	Jun	0.4259259
7	Jul	0.3620072
8	Aug	0.2939068
9	Sep	0.3333333
10	Oct	0.2939068
11	Nov	0.2666667
12	Dec	0.3548387

What patterns are you seeing in this additional data you are including in the model? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The above graph shows the Average fog days for each month of the year. We can see an increasing pattern of average fog from January to June, with June being the month of the highest fog days with 0.4259259.

Additionally, a decreasing trend can be seen from June to August and then a sudden increase in the month of September. Then, from September to November, there is a decrease in the average fog and finally a significant increase in the month of December.

Ice Pellets



	Month	Ice
1	Jan	0.003584229
2	Feb	0.003921569
3	Mar	0.000000000
4	Apr	0.000000000
5	May	0.000000000
6	Jun	0.000000000
7	Jul	0.000000000
8	Aug	0.000000000
9	Sep	0.000000000
10	Oct	0.000000000
11	Nov	0.000000000
12	Dec	0.007168459

What patterns are you seeing in this additional data you are including in the model? [Please note anything unusual or unexpected and describe the general trend or pattern of the data.]

The above graph shows the Average Ice pellet days for each month of the year. We can clearly see that the ice pellets are only visible in three months, with December being the highest ice pellets days month with 0.007168459. The months of January and February also show ice pellet days with 0.003584229 and 0.0039215669, however, these value of ice pellets is insignificant in front of December.

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

Unfavorable weather conditions can significantly affect how quickly an ambulance can get to its destination. High wind speeds, heavy precipitation, snowfall, and fog or ice pellets can generate dangerous driving conditions that could cause delays in ambulance response times.

Unfavorable weather might affect how accessible highways are. If roads aren't appropriately cleared after heavy snowfall, they may become inaccessible, and ice particles can make them dangerously slick. Due to the necessity for slower driving and more caution in such situations, ambulance response times may increase.

Ambulance drivers may encounter difficulties maneuvering safely due to fog and decreased visibility. Drivers may experience delayed reaction times due to having to drive more cautiously to prevent collisions.

Emergency call volume can also be influenced by weather conditions. A larger call volume and lengthier response times may result from increased accidents or health-related situations during severe weather.

To study the correlation between weather and response times, historical weather data and ambulance response time data should be gathered. Finding patterns and correlations through data analysis and modeling can assist in making informed decisions about how best to allocate resources.