GEORGE MASON UNIVERSITY

Data Analytics Engineering

FALL 2021

Report Team Smart Workers

DAEN 690 Project Report

TEAM SMART WORKERS

Yashashree Panda

Deepa Kapse

Harshita Thurlapati

Suraj Reddy Alimineti

Vaishnavi Rajput

George Mason University

George Mason University

DAEN 690 Capstone

Fall 2021

Fairfax County Fire and Rescue Department

Final Report

TEAM SMART WORKERS

Deepa Kapse (Product Owner & Developer)

Yashashree Panda (Scrum Master)

Suraj Reddy Alimineti (Developer)

Harshita Thurlapati (Developer)

Vaishnavi Rajput (Developer)

Under the guidance of

Professor F Brett Berlin

Table of Contents

Contents

Abstract

The Fire and Rescue Department of Fairfax (FRD) works around the clock to make sure the patients reach the right hospital on time, to achieve this FRD has a set of transport protocols that the providers are supposed to follow for transport decision making. Right now, FRD does not have any mechanism to programmatically evaluate the transport decision system. The FRD desires to visualize and find anomalies in transport decisions as they currently do not have a purposeful understanding of what are the factors influencing these decisions. The data was collected by FRD from incidents that took place in Fairfax from 2018 to 2020, it contains information about the hospitals, incidents, units, and patients. We built several classification models that continuously monitor patient, incident, unit, facility data to predict the probability of which facility patient should be taken in case of an incident. We drew conclusions from model predictions to find anomalies and further evaluate how various attributes contributed to these decisions. The model probability (confidence threshold) and mismatch percentage were considered to narrow down the anomalies from the dataset. We focused on the features which had a large percentage of mismatch and we found interesting patterns using meaningful visualizations in tableau. The visualizations permit the FRD to identify trends and realize potential incorrect decisions and related influential factors.

1. Introduction

1.1 Overview:

The Fire and Rescue Department of Fairfax (FRD) serves over 50,000 patients per year and makes sure they reach the hospital. Whenever any emergency occurs, the transport personnel must decide on where to take the patient. The start point of the incident is where the incident took place, and the endpoint of the incident is the hospital. The fire department personnel's work in three shifts in all the fire stations in Fairfax, so whenever any incident occurs, the personnel's go to the incident location and take victims to hospitals. FRD has a set of steps which they follow in every case, which are mentioned in Emergency Medical Dispatch cards. The steps have always been followed as a part of procedure, but the department never evaluated the system as a whole. The main aim of this project is to develop a mechanism to evaluate the system programmatically and automatically to determine if the steps taken were efficient or there was a scope of improvement. The main motive is to find out the anomalies in data to see if there is a pattern that is leading to wrong transport decisions so that FRD can address that pattern and respective measures can be taken to improve the transport decision. The FRD provided us with data which was collected over the years 2018, 2019 and 2020. This data explains the incident in detail by providing the location, unit shift time and tour on which the incident happened and, type of incident. The data also has details about the units and patients in detail. The medications and procedures which were used when the victims were taken to hospitals. We were also provided with a detailed list of all the facilities in and around Fairfax. This dataset has latitude, longitude data of the facilities and the designation lists of facilities which basically tells us what the facility treats. This all data combined will give us the whereabouts of an incident in detail. We will know the type of incident, which day of shift it was and which tour of shift was working. We will know which Unit responded to the incident which can give us the approximate or nearest possible location because the nearest unit always responds to incidents, we will also know the type of unit which responded to the incident, what primary actions units took in order to provide first aid to the victims. We will also know the level of care provided by units in incidents, the age of patients involved in the incident, the hospital they were taken to, the medications provided by the facility to victims and the procedures which were performed on victims in facilities. These all datasets combined gives us a detailed idea of what happened and how it was taken care of. And by analyzing this data combined, we are expected to find the incidents which could have been taken care of in a better way and if there is any pattern in such anomalies.

1.2 Problem Space:

In order to find out the solution of the problem, we have firstly performed the data quality assessment. Where we have assessed the completeness of the data, checked for the unique records in the dataset, looked for the lack of corrupt data in the dataset, and then acquired the accuracy in the dataset. The next step performed was data conditioning where we cleaned the data using different python libraries. And also handled the null values in this step. Later we got our data ready by performing data preprocessing. At this step we have done initial analysis to the dataset given to find out the most feasible hospital to which the patient can be taken when any incident occurs. Machine learning models such as Logistic Regression, Linear Regression, Random Forest, Decision tree and Naive Bayesian were being introduced. Finally, interactive dashboards were created which helped derive meaningful insights to improve the transport decision making and find the anomalies in the existing system.

1.3 Research:

The decision on where to take the injured person is made by the fire and rescue department depending on many factors. Some factors that were recognized are as follows:
The most important factor is where the patient wants to go when an injury occurs. The injured person mentions his/her preference then the fire and rescue department people have very little role in deciding. This is because when the injured person says that I want to go to a specific hospital or urgent care then the fire and rescue department must take him/her there.

The second most important factor is what kind of injury took place. It is important because it will filter out the destination hospitals more. For example, if an incident involves burns then the destination hospital must have facilities to treat burns. If the destination hospital can't treat burns, then there is no point in taking a patient there who is suffering from burns. Another example is if the patient is suffering trauma, then it is a critical case, and the person should be taken to a place where they can treat him/her.

The other influential factor is how far is the home hospital. If a person wants to be treated at home, then there are many questions which should be asked, like how far is the injured person's home? If the injured person stays an hour far then it doesn't make more sense to go home, rather it would be better to find a hospital nearby. Another question which should be asked is how much is the person injured? If the person is injured majorly and is insisting on taking him home, it is a very risky thing. The fire department should take him/her to the nearest possible hospital.

The last vital factor is the availability in the nearest hospital. Especially during these COVID times, the capacity of any hospital is a very important factor. If an incident occurs and the Fire and rescue department reaches that place and figures out a hospital nearby, they also must think and confirm if there is availability in that hospital or not.

1.4 Solution Space:

As accidents are a part of everybody's life, anyone and everyone can run into an accident unknowingly, because we all are humans. The process included to rescue and save a person after the accident has occurred is what we are working on. And it is important because we are not only trying to identify the flaws in the process, which is followed now but also, we are trying to find the reason behind those flaws so that we can analyze and rectify the process. This can save lives and reduce a person's suffering. The process in which we conducted our project goes as follows. The first step is data acquisition in which we had to collect the data given by our mentors. There are two data sets: Facilities, Individual data set in which there are multiple sheets of patients, units, incidents, etc. details. The next step is data quality assessment in which the quality of data is tested in the factors of completeness, uniqueness, integrity and accuracy. The third step is data conditioning. In this step we cleaned the data so that it is flexible for us to implement various kinds of models on it. Also, as a process of this step we had to also deal with the null values in the data. The following step is data preprocessing in which we found out the distances between each of the units to each of the facilities using their latitude, longitude coordinates and stored it as a dataset which further helped us with the project. The fifth step is to perform data analysis, and this is the most crucial stage of the project. In this we performed linear, logistic regression, random forest, decision tree and naive bayes. The final step is when we create data visualizations to communicate

our observations to the clients visually. We have created multiple visualizations and dashboards in this step.

1.5 Project Objectives:

As there are numerous employees working in the Fire and Rescue Department of Fairfax, they work in shifts all over the units and fire stations. We are working on this project to find out or identify the correlations between the shifts, units, hospitals nearby, type of injuries that occurred, to estimate what variables are reasonable for the decision of where to take the injured person? This data will in turn help us with the basic process of transport decision-making. The answers to all the questions above will in conclusion affect the way a victim will be transported to the desired suitable facility.

1.6 Primary User Story:

If a Fairfax County personnel is assigned to provide care to a patient, how should the personnel decide upon the facility to which the patient has to be sent. Many factors count in this such as the type of injury if it is physical or is trauma. The distance of the facility from the point of the incident also has to be counted. The medical history of the individual and their personal preference is also given importance. The main objective is to provide the best care to the patients.

1.7 Product Vision - Sample scenarios (why would someone want to use this):

The Fairfax County department wants to understand the factors that majorly influence the decision of finding a facility and transporting the patients there given all the factors of the situation. The distance from the incident to the facility, position of the patient, kind of injury from which they are suffering are considered for this. The basic product vision is that in case of emergencies the patient finds the best care possible, and the care provider can make decisions based on all the influential attributes and to what extent they make a difference.

❖ Scenario #1

When the patient makes a call, he/she might not be able to convey everything as he/she might be in pain. The health worker provider should be able to note down whatever the patient is going through as soon as possible once he reaches the incident place.

❖ Scenario #2

If in case of an accident the choice of facility the patient is showing priority to does not meet the requirements of the situation or does not make it possible for the caregiver to transport the patient there, the personnel must be incapacity to make decisions based on the other factors.

1.8 Definition of Terms:

EMS: Emergency Medical services.
EMD: Emergency Medical Dispatch.
ALS: Advanced Level Support.
BLS: Basic Level Support.
FRD: Fire and Rescue Department.
RF: Random Forest.
DT: Decision Tree.
LR: Logistic Regression.

LOC: Level of Care.
Lat: Latitude.
Long: Longitude.
PCR: Patient Care Report.
ML: Machine Learning.
EMS: Emergency Medical Services.
EPCR: Electronic Patient Care Record.
ePCRS: Electronic Patient Care Reporting System.
EDA: Exploratory Data Analysis46

2. Data Acquisition

2.1 Overview:

To start with our project, we have been provided with 2 datasets, one is the Facilities dataset and the other is the Individual Tables dataset. Explanation of both the datasets is as follows:

Facilities Dataset:

This dataset provides us with the knowledge of the hospital facilities available. This includes the name, address, and type of facility centers. We have details of a total of 32 hospital facilities, and it has 7 features.

Individual Tables Dataset:

We have been assigned a new dataset by our partners named Individual tables. This dataset has five sheets namely the incidents sheet, units' sheet, patients' sheet, procedures sheet, and medications sheet. The incidents sheet gives a complete picture of the incidents that took place and has attributes such as the primary key of the incident, incident type, incident description, and others. The unit sheet focuses on the fire station units it includes which unit has responded to the call, which unit has transported the vehicle to the place of incident, which unit type has gone for transportation. The patient's sheet gives us information about the patients for example, which unit was responsible to carry the patient, which hospital facility was the patient taken to, the patient's age, and the primary and secondary information about the patient. The procedures sheet depicts the procedures which were performed on the patient and the advanced life support given to the patient. The medications sheet gives us information about the medications that were given to the patient. We have also used other datasets from the Fairfax Fire Department website in order to get the address of each fire station.

2.2 Field Descriptions:

Facilities Data:

Facility_Name: Name of the facility
Facility_Location_Code: This field is a unique key used to identify the facility.
Facility_Full_Address: Address of facility.
Facility_Type_Of_Facility: The type of facility.
Facility_Latitude: Latitude coordinates of the facility.
Facility_Longitude: Longitude coordinates of the facility.
Facility_Hospital_Designation_List: Type of care facility is designed to offer.

Individual Tables Data:

- ***Incidents Sheet***
  PrimaryKey: Unique key to identifying the incident.
  CallConfirmedDT: DateTime stamp when the call comes to 911.
  ShiftDay: Which shift day was it A, B, C.
  TourOfShift: Day number of shift, each shift has 3 days in the span of 9 days.
  IncidentFirstDue: First verified the location of the incident, NULL if its outside Fairfax.
  InitialIncidentType: Type of incident identified when the call comes to 911.
  DispatchedIncidentType: Type of incident identified when a unit was dispatched.
  ArrivedIncidentType: Type of incident when the unit arrived there.
  FinalIncidentType: Final incident type when the call was closed.
  FinalIncidentTypeDescription: Description of the incident when the call was closed.
- ***Units Sheet***
  PrimaryKey: Unique key to identifying the incident.
  CallConfirmedDT: DateTime stamp when the call comes to 911.
  ResponseUnitID: ID of the unit which responded to the incident.
  TransportUnitID: ID of the unit which transported the patient.
  TransportUnitLOC: Level of Care of transporting units.
  UnitPrimaryActionTaken: Primary action was taken by unit, recorded an ePCR.
  UnitOtherActionsTaken: Other actions taken by a unit.
  UnitStation: Fire station to which the unit is assigned.
  UnitType: Type of Unit
  UnitArrivalOrder: The order in which the unit arrived at the incident location.
  UnitTransportedDT: The date time stamp of the unit leaving the incident location.
  UnitTransportedArrivalDT: The date time stamp of the unit reaching the hospital.
- ***Patients Sheet***
  PrimaryKey: Unique key to identifying the incident.
  CallConfirmedDT: DateTime stamp when the call comes to 911.
  PatientID: Unique number to identify the patient.
  TransportUnitID: ID of the unit which transported the patient.
  TransportLOC: Level of Care of transporting units.
  TransportHospitalCode: Unique number of Destination of the transport(Hospital).
  TransportHospitalName: Name of the hospital.
  ReasonForChoosingHospital: Reason to choose that particular hospital.
  PatientAge: Age of the patient.
  PrimaryImpression: Primary clinical impression of the patient indicated by the provider.
  SecondaryImpression: Secondary impression of the patient indicated by the provider.
  PatientDisposition: Outcome from the patient regarding FRD.
- ***Procedures Sheet***
  PatientID: Unique number to identify the patient.
  ProcedurePrimaryKey: Unique number to identify the procedures performed.
  ProcedurePerformed: Procedures performed on patients.
- ***Medications Sheet***
  PatientID: Unique number to identify the patient.
  MedicationPrimaryKey: Unique number to identify medication administration.
  MedicationGiven: Name of medications given.

2.3 Data Context:

The Fairfax Fire and Rescue Department works in challenging times ensuring the health and safety of residents of Fairfax as well as their employees. They work on five core values, **P**rofessionalism, **R**espect, **I**ntegrity, **D**iversity, **E**xcellence, and they work to save our lives. The Individual tables data contains details of 4 main things. First are the unit details, the shift and tour of shift details of the unit, and all other unit information. The second thing is the incident information, location, and type of incident. The third thing which has detailed fields in the dataset is the patient, and all the patient information is provided in the dataset. The last thing is procedures, when the patient is taken to the hospital, some procedures are performed on the patient, which is explained in the dataset. Another piece of data which we have is the list and functional details of Facilities, which is basically all the details of the hospitals around the city. It includes details of the name, location of the facility as well as the service which the facility is meant to provide.

2.4 Data Quality Assessment:

There is a presence of null values in almost 21 features of the dataset. Out of which IncidentFirstDue feature has the significance of null values. It shows that the incident was outside of Fairfax. Other than that, features which have significant amounts of null values are UnitArrivalOrder, ProcedurePrimaryKey, ProcedurePerformed, MedicationPrimaryKey, MedicationGiven, UnitTransportedDT, UnitTransportedArrivalDT, UnitOtherActionsTaken. Within these eight features, there are 54000 null values on average. These values should be handled in a way that the dataset makes sense. We are checking with our partners if we can use any other columns to fill these null values or if they can provide us with additional resources to fill this gap. The two fields with more null values are date stamp fields: UnitTransportedDT, UnitTransportedArrivalDT fields, which need to be handled in a correct way so that the data makes sense. We have two datasets as mentioned earlier -The Individual Tables dataset and the Facilities dataset. This table below gives details such as completeness, uniqueness, integrity, accuracy, consistency, etc. which define the quality of the data we have overall.

Quality	**Facilities Dataset**	**Individuals Dataset**
Completeness	94%	92%
Uniqueness	✅	✅
Integrity	✅	✅
Accuracy	✅	✅

Number of attributes with null values	4	21
Attributes to missing values	Attributes of Facility Data set with Missing value: - 4 Attributes	Attributes of Individuals Dataset with Missing value: - 21 Attributes
Consistency	✅	✅
Conformity	✅	✅

Figure 1 *Data quality assessment*

2.5 Data Conditioning:

2.5.1 Handling Null Values

We have handled null values in various fields in different ways, as the fields were of different types, each one of them needed a different approach to handle. Following is the explanation for each field and how we handled null values in it.

Feature – UnitOtherActionsTaken: There is a total of 61.64% of null values in this column. This column represents the actions other than primary actions taken by a unit. There is a need for the replacement of all these null values, the most relevant value for this feature for replacement we found is "no other action". This value in the column says that other than primary action no other action was taken by the unit, which makes sense with the null values because that can be the reason it was left null.

Feature – UnitTransportedArrivalDT: There is a total of 57.6% of null values in this column. This column indicates the date-time stamp of the unit when they reach the hospital.

Feature – UnitTransportedDT: There are a total of 57.13% of null values in this column. This column indicates the date-time stamp of the unit when they reach the hospital.

Feature – MedicationGiven: This column has around 45% of null values in the column. This column represents the medication given by the hospital to the patient. The top 5 medications given in the dataset over a period of 3 years are Epinephrine 0.1 MG/ML (1:10,000), Oxygen, Normal saline, Fentanyl Citrate (Sublimaze), Ondansetron (Zofran)

Feature – ProceduresPerformed: This column has around 21.62% of null values in the column. This column represents the procedures performed by the hospital to the patient. The top 3 procedures performed in the dataset over a period of 3 years are IV Start - Extremity Vein (arm or leg), CV - ECG - 12 Lead Obtained, CV - Defibrillation - Manual

2.5.2 Data Cleaning

As we plan to go further with the project, we are close to start applying algorithms and starting analysis on the data we have. But before starting to apply algorithms we must prepare the data to be compatible with any algorithms. The data has many null values as we discussed in the previous part. We had to find a way to handle them accurately. After discussing with the partners, we decided to handle them in different ways which will be discussed in this section. As we have two datasets that we work on, we have handled the null values of both datasets differently.

Facilities dataset

a. Feature - Facility_Hospital_Designation_List

For this feature we had some null values initially, this feature is used to indicate what is that hospital special for, for example, some hospitals have specialty for burns, some have for trauma. Initially, we had null values for 5 types of records.

Referred hospital description for filling missing values

Facility_Name	Facility_Hospital_Designation_List	Check with partners
Landing Zone / Non-Hospital	non hospital	Confirmed
Kaiser Permanente - Tysons Corner - Advanced U...	Hospital General	Confirmed
MedStar Southern Maryland Hospital Center	Hospital(General), Burns, Stroke, trauma	Confirmed
HCA StoneSprings Hospital Center	Hospital General	Confirmed
Out Of Area Hospital Not Listed	Hospital General	Confirmed

Figure 2 *Hospital description for filling the missing values*

We had a discussion with Fairfax Fire and Rescue Department they suggested us with the following replacements

- Kaiser Permanente - Tysons Corner - Advanced Urgent Care, Kaiser Permanente - Woodbridge - Advanced Urgent Care, HCA StoneSprings Hospital Center, Out of Area Hospital Not Listed - should be replaced with Hospital (General)
- Landing Zone / Non-Hospital, Landing Zone - should be replaced with non-Hospital
- MedStar Southern Maryland Hospital Center - should be replaced with Hospital (General), Burns, Stroke, trauma.

b. Feature - Facility_Type_Of_Facility

This feature indicates the type of hospital, for example, some values in the feature are Hospital, Emergency room. There were some null values in this feature which were replaced with ‘non-Hospital service’.

Individual Tables Dataset

After observing the dataset which we were given, we understood there are empty values in the dataset. Our first observation was that there were very few missing values, so we have dropped the

records of columns named IncidentFirstDue, ArrivedIncidentType, FinalIncidentTypeDescription.

	Dtype	Count of missing	% of missing
PrimaryKey	int64	0	0.000000
CallConfirmedDT	datetime64[ns]	0	0.000000
ShiftDay	object	0	0.000000
TourOfShift	int64	0	0.000000
IncidentFirstDue	float64	14	0.000114
InitialIncidentType	object	0	0.000000
DispatchedIncidentType	object	0	0.000000
ArrivedIncidentType	object	389	0.003175
FinalIncidentType	object	0	0.000000
FinalIncidentTypeDescription	object	42	0.000343

Figure 3 *Attributes in the dataset corresponding to their data types and count of null values*

In the unit's sheet, we observed there are missing values in some of the columns and we are displaying the missing values columns below.
In the patient's sheet, we have observed that there are a few null values, so we have replaced all the null values of the 'TransportHospitalName' column with 'No Hospital service'. Then we replaced all the null values of the 'TransportHospitalCode' column with 'No Hospital name'. After that, we replaced all the null values of the 'ReasonForChoosingHospital' column with 'No Hospital service'. And replaced all the null values of the 'TransportLOC' column with '0'. There were no missing values in the Procedures and medications sheets.

2.5.3 Data Preprocessing

Step 1: Cleaning the dataset & dealing with different features

a. Dropped the nominal values with large variability. Features such as UnitTransportedArrivalDT, PrimaryKey, PatientID, UnitTransportedArrivalDT, 'UnitOtherActionsTaken','SecondaryImpression', CallConfirmedDT, had large variability within the data around 40-70% of the data values were unique.
b. Dropped highly correlated features. There were multiple features dealing with the same set of information. For example, the features related to incident type category such as InitialIncidentType, DispatchedIncidentType, ArrivedIncidentType, FinalIncidentType, FinalIncidentTypeDescription dealt with related set of information. We chose to keep one or two important features which represented the entire category.
c. Calculated distance between each Unit ID and Facility using haversine formula. Included distance between each UnitID and facility using haversine formula. Wrote a python function to calculate the distance and stored it in a csv file for further use in modeling.

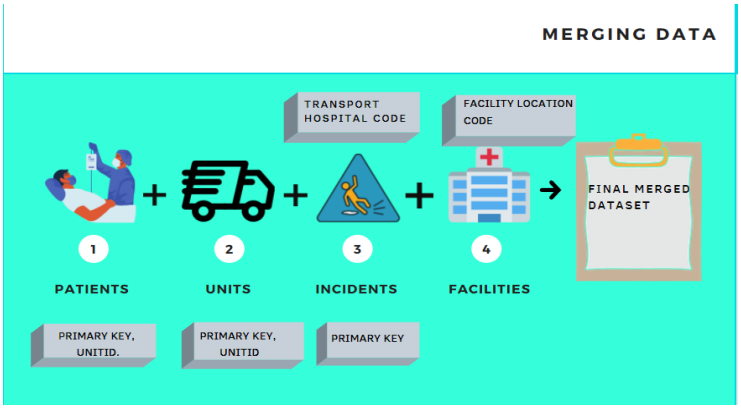
Step 2: Merging different datasets.

a. Merge patients with the unit's dataset. Each 'Units' had 0: many 'Patients'. Merged Patients with the unit's dataset using features "PrimaryKey", "TransportUnitID", "ResponseUnitID". A new merged "PatientUnitsMerge" dataset was created.
PatientUnitsMerge = patients.merge(units, left_on=["PrimaryKey", "TransportUnitID"], right_on= ["PrimaryKey", "ResponseUnitID"])
b. Merge PatientsUnits with the incident's dataset. Merged "PatientUnitsMerge" with the incident's dataset using features "PrimaryKey". A new merged "IncidentPatientUnitsMerge" dataset was created.

c. Merge IncidentPatientUnitsMerge with the facilities dataset. Each 'Incident' had 0: many 'Patients' & 'Units'. Merged "IncidentPatientUnitsMerge" with the facilities dataset using features "TransportHospitalCode" and "Facility_Location_Code". A new merged "IncidentPatientUnitsFacilitiesMerge" dataset was created.

IncidentPatientUnitsFacilitiesMerge = IncidentPatientUnitsMerge.merge(facilities, how="left", left_on="TransportHospitalCode", right_on="Facility_Location_Code")

The "IncidentPatientUnitsFacilitiesMerge" merged dataset was further transformed and used for building models.



Merging Dataset

Step 3: Converting categorical values into numerical values.

a. Since we were planning to use a classification model converting categorical features into numerical values was a must.
b. We chose to use pandas get_dummies () for converting categorical into numerical values.

2.6 Other Data Sources:

EMD Cards – We were provided with the EMD Cards from FRD. Which has a flowchart which is followed whenever a call is received by 911. And then the EMD cards are filled in based on the information FRD has.
The sections in EMD cards are as follows
a. Vital points questions: These are the questions asked related to the patient's condition which is understood, questions like are the person having chest pain? Did the person pass out?
b. ALS Priority Response: Questions related to the response needed from the Advance Level Support Unit.
c. BLS Priority Response: Questions related to the response needed from the Basic Level Support Unit.
d. Pre-Arrival Instructions: These are the things that are used as guidance given to patients before the unit is reached there.
e. Short Report: This is a brief explanation about the condition of the patient, and surroundings.
f. Background Information: This is a section that provides knowledge about the type of injury, it explains in detail how and why that injury can be caused.

3. Analytics and Algorithms

From the latitude and longitude data information provided by the partners, we determined the distances between each of the units and the facility hospital. We calculated the distance in Excel using the formula below.

- ACOS [(sin (Lat_place_1*PI ()/180) *sin (Lat_place_2*PI ()/180) + cos (Lat_place_1*PI ()/180) *cos (Lat_place_2*PI ()/180) *
- cos (Lon_place_2*PI ()/180-Lon_place_1*PI ()/180))] *3443.8985

		LAT	38.4239	38.9291	38.4153	38.6384	38.8635	39.0003	38.9977	38.7406	38.748	38.995	38.9272	38.9371
		LON	77.4016	77.2246	77.4075	77.294	77.2342	77.0949	77.11	77.0774	76.8765	77.4805	77.0145	77.109
			Novant He	Kaiser Per	MWHC St	Kaiser Per	Merrifield	Walter Re	Suburban	Inova Mo	MedStar S	Inova Hea	Children's	Sibley Me
LAT	LON	FACILITY_NAME_LABEL												
38.9317	77.1776	Fire Station 1 - McLean	37.112	2.53021	37.7805	21.2149	5.60933	6.50475	5.83195	14.2652	20.5894	16.8541	8.77402	3.70709
38.8991	77.2615	Fire Station 2 - Vienna	33.6999	2.86922	34.3542	18.1037	2.86611	11.3621	10.613	14.7718	23.2088	13.5114	13.4177	8.60557
38.8436	77.3056	Fire Station 3 - City of Fairfax	29.4654	7.34228	30.1106	14.1985	4.07996	15.6741	14.9638	14.205	24.039	14.0662	16.688	12.3914
38.9691	77.3859	Fire Station 4 - Herndon	37.6854	9.10064	38.2921	23.3852	10.9475	15.7849	14.9546	22.9142	31.3863	5.38758	20.1707	15.0444
38.7807	77.1464	Fire Station 5 - Franconia	28.2477	11.0911	28.9238	12.6519	7.42441	15.4318	15.1226	4.63343	14.7149	23.292	12.3621	10.9978
38.8325	77.1917	Fire Station 8 - Annandale	30.4278	6.90669	31.1006	14.5056	3.13059	12.7128	12.2283	8.84805	17.9567	19.1632	11.5599	8.48743
38.7428	77.0774	Fire Station 9 - Mount Vernon	28.1506	15.1231	28.8155	13.7326	11.8717	17.8251	17.7019	0.15211	10.8311	27.8283	13.1842	13.5398
38.8455	77.1393	Fire Station 10 - Bailey's Crossroads	32.3935	7.38022	33.0694	16.564	5.25621	10.9651	10.6333	7.97987	15.6773	21.0559	8.76969	6.53915
38.773	77.0794	Fire Station 11 - Penn Daw	29.7488	13.3188	30.4186	14.8485	10.4159	15.7295	15.6088	2.24716	11.0721	26.4725	11.2089	11.4507
38.9986	77.2909	Fire Station 12 - Great Falls	40.164	5.98108	40.8032	24.8972	9.82635	10.531	9.71793	21.2138	28.2361	10.1872	15.6516	10.6574
38.9025	77.2237	Fire Station 13 - Dunn Loring	34.4416	1.84007	35.1044	18.6434	2.75785	9.67544	8.97449	13.6839	21.5273	15.213	11.373	6.61066
38.7932	77.2719	Fire Station 14 - Burke	26.4647	9.7332	27.1255	10.7657	5.26342	17.1934	16.598	11.0938	21.5341	17.8974	16.6603	13.2574
38.8919	77.4316	Fire Station 15 - Chantilly	32.3828	11.4236	32.9668	19.0255	10.7988	19.5874	18.7656	21.749	31.4987	7.5929	22.5581	17.6223
38.7809	77.3848	Fire Station 16 - Clifton	24.688	13.3884	25.2994	11	9.91358	21.752	21.0437	16.7952	27.4827	15.6646	22.3424	18.3504
38.837	77.4299	Fire Station 17 - Centreville	28.5867	12.7498	29.1698	15.5584	10.6909	21.2603	20.4744	20.1218	30.4366	11.253	23.1985	18.5964
38.8658	77.1928	Fire Station 18 - Jefferson	32.5545	4.69588	33.225	16.6384	2.23297	10.6826	10.1389	10.657	18.8824	17.8574	10.4835	6.67606
38.7034	77.2106	Fire Station 19 - Lorton	21.8989	15.62	22.5752	6.35738	11.1364	21.4469	21.0453	7.62601	18.2737	24.8429	18.7253	17.0546

Figure 4 *Distances calculated from every unit to every facility*

Analytics:

We have analyzed the given data and figured out that we need to know the greatest number of procedures performed. From the below bar chart, we can see that the Procedures performed are given on X-axis and the number of procedures performed is given on Y-axis. The medical procedure named "CV- ECG- 12 Lead Obtained" was the most common procedure performed followed by the procedure named "IV Start - Extremity Vein (arm or leg)".

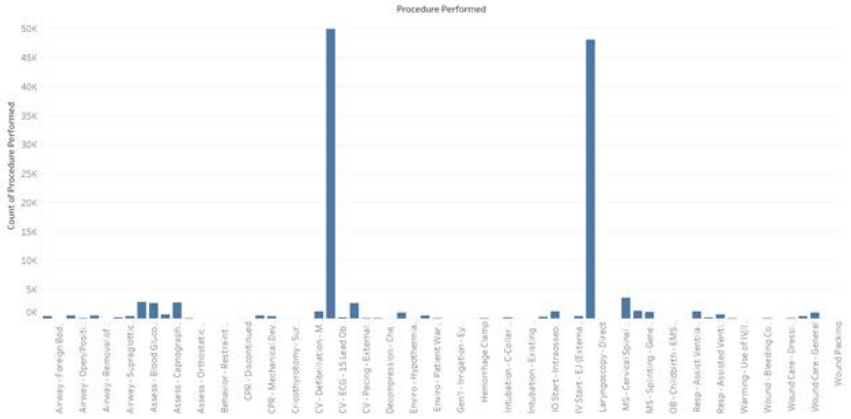


Figure 5 *Visualization describing the count of each Procedure Performed*

From the graph below, we can see that the medications given to the patient are given on Y-axis and the count of medications given is given on X-axis. And the most common medication given to the patients is named "Oxygen" followed by the medication "Normal saline".

Observed Analysis for building models:
As we are trying to find out the distance between the transport unit and facility. We considered the associated label for each facility fire station name from the fire stations dataset and created a new sheet in excel to place these attributes. From the Individual table dataset, we have considered the incident primary key, and the transport hospital associated with each incident primary key. As mentioned earlier we have tried to find out the distance between each facility hospital and fire station unit. So now we are trying to find out the distances between each fire station and facility hospital associated with each incident. This will give a better idea of how much time each unit is taking to complete the task. So, with this analysis, when we come across a new incident, we can use this model to look for the nearest possible five hospital names and detect the outliers involved which affect the decision on choosing the hospital.

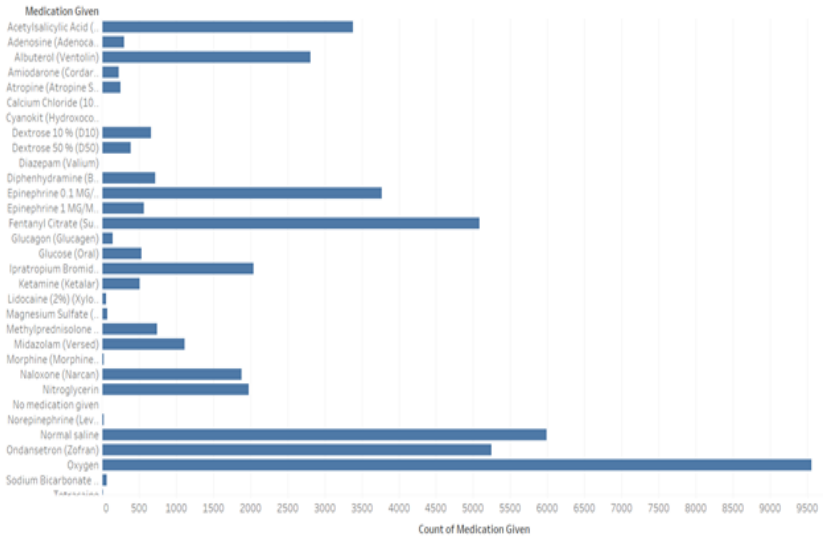


Figure 6 *Visualization describing the count of each medication given*

Logistic Regression:
In order to find which hospital, the patient should be taken to, from the place of the incident we have analyzed the data and found out the influential variables. These influential variables are "TransportUnitID" and "TransportLOC" which we used from the Patients sheet. Below is the screenshot of the sheet which we used.

PrimaryKey	CallConfirmedDT	PatientID	TransportUnitID	TransportLOC	TransportHospitalCode	TransportHospitalName	ReasonForChoosingHospital
9804225142	2018-07-01 00:09:19	640779	M422	Level 4: 1 Provider (BLS Care)	205	Springfield Health Plex	Closest Facility
9804225144	2018-07-01 00:09:24	640780	M435	Level 4: 1 Provider (BLS Care)	205	Springfield Health Plex	Closest Facility
9804095125	2018-07-01 00:05:40	640786	M409B	Level 4: 1 Provider (BLS Care)	47	Mount Vernon Hospital	Closest Facility
9804105181	2018-07-01 00:25:54	640790	M410B	Level 3: 1 Provider (ALS Care)	16	Fairfax Hospital	Patient's Choice
9804095125	2018-07-01 00:05:40	640792	M409	Level 2: 2 Providers (ALS Care)	47	Mount Vernon Hospital	Protocol,Specialty Center (Trauma, STEMI, Stroke)
9804085219	2018-07-01 00:47:45	640803	M408	Level 3: 1 Provider (ALS Care)	16	Fairfax Hospital	Patient's Choice
9804105281	2018-07-01 01:21:15	640806	M428	Level 4: 1 Provider (BLS Care)	3	Virginia Hospital Center	Closest Facility

Figure 7 *Influential attributes from the Patients sheet used for the Logistic Regression model*

Then we have checked for the unique values in each of the columns. And we have chosen TransportUnitID, TransportLOC and TransportHospitalName. Where TransportUnitID and TransportLOC are x-axis and TransportHospitalName is on the y-axis as required for Logistic Regression. To perform Logistic Regression, we needed integer type variables, so we have converted all the object type variables into integer type variables. As TransportHospitalName has a categorical data type in it, so we have assigned each of the Hospital names into a specific code using Python libraries. And we have changed the "No Hospital Service" to 111. Later, we divided the data into test and train datasets to build the Logistic Regression model. After splitting the data, we performed a Logistic Regression model which gave us an accuracy of 72%. Based on these two variables "TransportUnitID", "TransportLOC" we will be able to predict which hospital the patient has to be taken to with 72 percent accuracy. By adding parameters, we can increase the accuracy of the predictions.

```
accuracy = total_correct/len(y_test)
print(accuracy)
```

```
0.7202804660214882
```

Figure 8 *Accuracy of the Logistic Regression model*

Linear Regression:

The variables used to build the linear regression model are PrimaryKey ,ShiftDay ,TourOfShift ,IncidentFirstDue ,ResponseUnitID ,UnitStation ,UnitArrivalOrder ,TransportHospitalCode

```
Call:
lm(formula = Distance ~ PrimaryKey + ShiftDay + TourOfShift +
    IncidentFirstDue + ResponseUnitID + UnitStation + UnitArrivalOrder +
    TransportHospitalCode, data = my_data)

Residuals:
       Min         1Q     Median         3Q        Max
-1.374e-10 -1.600e-12 -2.000e-13  1.500e-12  2.015e-08

Coefficients:
                     Estimate Std. Error    t value Pr(>|t|)
(Intercept)          2.300e+01  3.432e-12  6.701e+12  < 2e-16 ***
PrimaryKey          -1.171e-21  9.282e-23 -1.262e+01  < 2e-16 ***
ShiftDay            -2.156e-14  5.504e-14 -3.920e-01    0.695
TourOfShift         -8.896e-14  8.449e-14 -1.053e+00    0.292
IncidentFirstDue    -2.888e-13  6.999e-15 -4.127e+01  < 2e-16 ***
ResponseUnitID       2.903e-14  5.227e-15  5.554e+00 2.79e-08 ***
UnitStation          1.392e-13  7.957e-15  1.750e+01  < 2e-16 ***
UnitArrivalOrder    -4.498e-14  3.709e-14 -1.213e+00    0.225
TransportHospitalCode  4.663e-16  1.069e-15  4.360e-01    0.663
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.731e-11 on 293471 degrees of freedom
  (311 observations deleted due to missingness)
Multiple R-squared:    0.5,     Adjusted R-squared:    0.5
F-statistic: 3.668e+04 on 8 and 293471 DF,  p-value: < 2.2e-16

>
```

Figure 9 *Linear Regression model summary*

From the model summary, we can see that Distance is most affected by the TransportHospitalCode which is clear from the estimated values.

Random Forest:

For Logistic Regression we considered three variables for the model building which are Transport UnitID, Transport LOC, and Transport Hospital code. Whereas for Random Forest we are

considering more numbers of attributes. As mentioned previously our new data has 5 separate spreadsheets each representing a different category. Firstly, we have merged the Patients and Units sheet where we have observed that there are now 288314 rows and 22 columns. After this, we have merged incident data with this which came to be 288314 rows and 31 columns.

```
In [111]: incidentpatientunitsmerge.head(15)
```

Out[111]:

	PrimaryKey	CallConfirmedDT_x	PatientID	TransportUnitID	TransportLOC	TransportHospitalCode	TransportHospitalName	ReasonForChoosingHospita
0	9804015371	2018-07-01 02:16:24	640838	M401	Level 3: 1 Provider (ALS Care)	3	Virginia Hospital Center	Patient's Choice,Closes Facilit
1	9804015371	2018-07-01 02:16:24	640838	M401	Level 3: 1 Provider (ALS Care)	3	Virginia Hospital Center	Patient's Choice,Closes Facilit
2	9804016046	2018-07-01 11:20:51	641032	M401	Level 4: 1 Provider (BLS Care)	3	Virginia Hospital Center	Patient's Choic
3	9804018610	2018-07-02 09:31:06	641839	M401	Level 4: 1 Provider (BLS Care)	16	Fairfax Hospital	Closest Facilit
4	9804018610	2018-07-02 09:31:06	641839	M401	Level 4: 1 Provider (BLS Care)	16	Fairfax Hospital	Closest Facilit
					Level 4: 1			Closest Facility Famil

Figure 10 *Merged data from patients' sheet, Unites sheet, and incident data*

We later performed feature engineering. In this step of the analysis, we had to check for categorical attributes in the current data. These categorical values were then converted in order to make it flexible to build the models we desire. We changed the datatypes of attributes 'TransportUnitID','TransportLOC','ReasonForChoosingHospital','PatientDisposition', 'TransportUnitLOC', 'ShiftDay', 'FinalIncidentType'. For the Unit station column, we observed that for some unit stations values were assigned as PRV rather than numeric, we replaced them with 111. Below is the screenshot of variables on which we performed feature engineering

```
Feature_list = ['TransportUnitID', 'TransportLOC', 'ReasonForChoosingHospital',
        'PatientDisposition', 'TransportUnitLOC',
       'ShiftDay', 'FinalIncidentType']
```

Figure 11 *Feature engineering*

After changing the data types, we dropped all the nominal features that are converted into numerical features using get_dummies. We divided the updated dataset into train and test datasets. The Ratio was 70:30. Then we implemented 4 Random Forest Models. From all 4 models built we observed that model 1 has the best performance with

```
Best n estimator: 25
Best max_depth: 15
Best min_samples_split: 2
```

Figure 12 *Performance of the best model in Random Forest*

And with accuracy for the train is 0.70 and with accuracy for test as 0.69.

	precision	recall	f1-score	support
1	0.95	0.01	0.02	10464
3	0.00	0.00	0.00	9218
11	0.87	0.66	0.75	26693
16	0.62	0.97	0.75	83927
36	0.00	0.00	0.00	243
47	0.86	0.85	0.86	23513
58	0.00	0.00	0.00	371
59	1.00	0.01	0.02	200
95	0.87	0.72	0.79	28849
102	0.00	0.00	0.00	196
103	0.00	0.00	0.00	1066
107	0.00	0.00	0.00	63
111	1.00	1.00	1.00	696
116	0.00	0.00	0.00	4
202	0.00	0.00	0.00	1123
204	0.00	0.00	0.00	229
205	1.00	0.00	0.00	7049
213	1.00	0.00	0.00	4865
215	0.00	0.00	0.00	705
216	0.00	0.00	0.00	16
324	0.00	0.00	0.00	135
325	1.00	0.00	0.01	414
326	0.00	0.00	0.00	124
327	1.00	0.02	0.04	176
328	0.00	0.00	0.00	96
363	0.00	0.00	0.00	12
364	0.00	0.00	0.00	15
374	0.00	0.00	0.00	714
700	0.00	0.00	0.00	4
701	0.00	0.00	0.00	3
999	1.00	0.06	0.11	115
accuracy			0.70	201298

Figure 13 *Summary of the Random Forest models*

Random Forest with all features

As suggested by the partners we transformed the TransportUnitID was trimmed down and the actual number part of TransportUnitID was used to build the model. Initially, the column had values shown in the following image

```
In [124]: incidentpatientunitsmerge.TransportUnitID.unique()
Out[124]: array(['M401', 'M402', 'M434', 'M404', 'M436', 'M405', 'M437', 'M426',
                 'M408', 'M408B', 'M430', 'M410B', 'M409B', 'M409', 'M424', 'M410',
                 'M428', 'M411B', 'M411', 'M412', 'M439', 'M413', 'M441', 'M432',
                 'M414', 'M427', 'M423', 'E415', 'M415', 'M421', 'M417', 'M416',
                 'M438', 'M418', 'M419', 'M420', 'M422', 'M440', 'M435', 'M425',
                 'M429', 'M442', 'M431', 'M401E', 'E401', 'A410E', 'M412E', 'A422E',
                 'A402E', 'M422B', 'EMS403', 'R418', 'A414E', 'A421E', 'E411',
                 'E439', 'A414', 'E418', 'E442', 'M405E', 'A402', 'E440', 'A438E',
                 'R411', 'TL436', 'M417E', 'A405E', 'E430', 'E423', 'TL440', 'E409',
                 'A413E', 'E402B', 'TL401', 'A417E', 'M414E', 'E428', 'R439',
                 'A401E', 'A417', 'HM440', 'R401', 'E405', 'A422', 'TL405', 'E417',
                 'M422E', 'A437E', 'R421', 'E402', 'E413', 'A410', 'E404', 'E421',
                 'E410', 'A401', 'A408E', 'M402B', 'A412E', 'E436', 'E432', 'T411',
                 'E434', 'E437', 'E438', 'TT430', 'E429', 'E420', 'R426', 'E426',
                 'E408', 'E414', 'M402E', 'A413', 'T430', 'M414B', 'EMS404',
                 'M401B', 'A422F', 'EMS405', 'E435', 'R414', 'TT410', 'A421',
                 'T422', 'M417B', 'EMS401', 'E422', 'TL438', 'E424'], dtype=object)
```

Figure 14 *Display of TransportUnitID array from the merged data*

And it was trimmed down to the following array

```
([401, 402, 434, 404, 436, 405, 437, 426, 408, 430, 410, 409, 424,
       428, 411, 412, 439, 413, 441, 432, 414, 427, 423, 415, 421, 417,
       416, 438, 418, 419, 420, 422, 440, 435, 425, 429, 442, 431, 403])
```

This means the Suffix and prefix of the values in the column were dropped and as there was no alphabetical part left in the column, we stored it as int. Then the getdummies () function was used to get dummies of some features and then the dummies were merged with the original data which is a combination of patient, incident, and unit datasets. And then the data was divided into X and Y and train and split, the Y variable is TransportHospitalCode and the rest of the dataset is stored in the X variable and the training dataset has 70% of the data vs the test dataset has 30%. We observed the following values as the best parameters

```
Best n estimator: 25
Best max_depth: 15
Best min_samples_split: 2
```

Figure 15 *Summary details of the best model*

We got the accuracy for the train as 0.58 and with accuracy for the test as 0.49. The feature Importance plot we got from this model is as follows

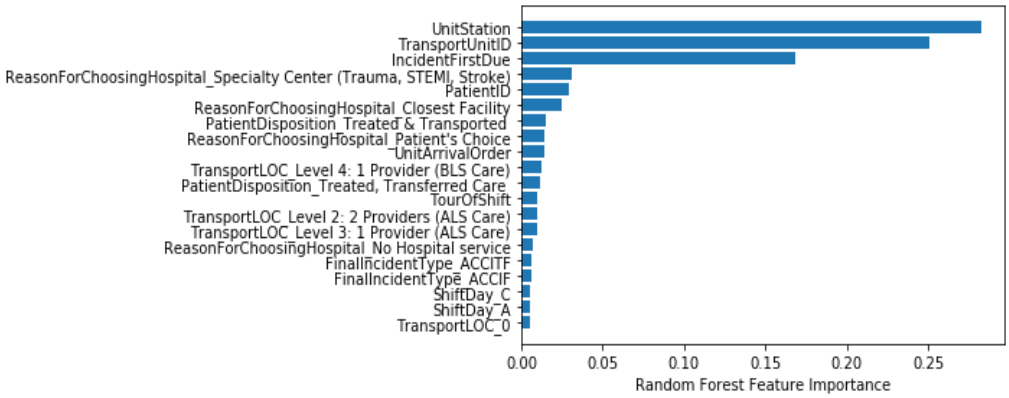


Figure 16 *Feature importance of Random Forest model*

	precision	recall	f1-score	support
1	0.79	0.02	0.04	4483
3	0.94	0.03	0.06	4010
11	0.67	0.28	0.39	11414
16	0.52	0.97	0.68	35883
36	0.00	0.00	0.00	91
47	0.75	0.65	0.70	10060
58	0.00	0.00	0.00	170
59	0.00	0.00	0.00	73
95	0.78	0.33	0.46	12442
102	1.00	0.04	0.07	85
103	0.00	0.00	0.00	485
107	1.00	0.38	0.55	16
111	1.00	1.00	1.00	298
116	0.00	0.00	0.00	2
202	1.00	0.00	0.00	506
204	0.00	0.00	0.00	85
205	0.96	0.01	0.02	3078
213	1.00	0.01	0.03	1976
215	1.00	0.00	0.01	319
216	0.00	0.00	0.00	1
324	1.00	0.12	0.21	59
325	1.00	0.35	0.52	189
326	0.00	0.00	0.00	54
327	1.00	0.26	0.41	70
328	0.00	0.00	0.00	40
363	0.00	0.00	0.00	3
364	0.00	0.00	0.00	5
374	0.00	0.00	0.00	319
700	0.00	0.00	0.00	2
999	1.00	0.09	0.17	53
accuracy			0.57	86271
macro avg	0.51	0.15	0.18	86271
weighted avg	0.67	0.57	0.50	86271

Figure 17 *Summary of all the iterations*

Random Forest with top 20 features:

The Random Forest model helped us to get the feature importance graph which shows that the top 20 features include UnitStation, TransportUnitID, IncidentFirstDue, ReasonForChoosing, TransportLOC, ShiftDay. We have used the same features to build a second random forest model which gave us the following best parameters and accuracies

```
Best n estimator: 25
Best max_depth: 15
Best min_samples_split: 2
```

Figure 18 *Summary of the best iteration in Random Forest when performed with the top 20 features*

The accuracy of the train dataset is 0.58 and the accuracy of the test dataset is 0.58

Decision Tree with all features:

The dataset we built in the last model with the feature engineering was used as it is for this model as well. The X variable formed in the last model with 70% of train data was used to build the model and then the model was used to predict the 30% test data. The accuracy of the training dataset is 0.56 and the F1 score is 0.49 and the accuracy of the test dataset is 0.56 and the F1 score is 0.49.

	precision	recall	f1-score	support
1	0.79	0.02	0.04	4483
3	0.94	0.03	0.06	4010
11	0.67	0.28	0.39	11414
16	0.52	0.97	0.68	35883
36	0.00	0.00	0.00	91
47	0.75	0.65	0.70	10060
58	0.00	0.00	0.00	170
59	0.00	0.00	0.00	73
95	0.78	0.33	0.46	12442
102	1.00	0.04	0.07	85
103	0.00	0.00	0.00	485
107	1.00	0.38	0.55	16
111	1.00	1.00	1.00	298
116	0.00	0.00	0.00	2
202	1.00	0.00	0.00	506
204	0.00	0.00	0.00	85
205	0.96	0.01	0.02	3078
213	1.00	0.01	0.03	1976
215	1.00	0.00	0.01	319
216	0.00	0.00	0.00	1
324	1.00	0.12	0.21	59
325	1.00	0.35	0.52	189
326	0.00	0.00	0.00	54
327	1.00	0.26	0.41	70
328	0.00	0.00	0.00	40
363	0.00	0.00	0.00	3
364	0.00	0.00	0.00	5
374	0.00	0.00	0.00	319
700	0.00	0.00	0.00	2
999	1.00	0.09	0.17	53
accuracy			0.57	86271
macro avg	0.51	0.15	0.18	86271
weighted avg	0.67	0.57	0.50	86271

Figure 19 *Summary statistics for the decision tree model with all features*

Decision Tree with top 20 features:

To estimate the best parameters for estimator model, I chose 'GridSearchCV' library from sklearn. It allows us to input multiple values for each parameter of decision tree algorithm and builds several models based on the various parameters. Also, it splits the data into 5 parts. It trains the model on first random 4 parts and cross validate on the 5th split of data. I can also add f1_score as a metric for algorithm evaluation.

Ran the estimator model for various parameters criterion = ['gini', 'entropy'], max_depth = [50,100,150,200], max_features = [100, 200, 250,300], max_leaf_nodes = [20,50,60,100,120].
The Decision tree model helped us to get the feature importance graph which shows that the top 20 features include UnitStation, TransportUnitID, IncidentFirstDue, ReasonForChoosing, TransportLOC, ShiftDay. We have used the same features to build a decision tree model which gave us the following accuracies

- ➔ The accuracy of the training dataset is 0.56 and the F1 score is 0.49.
- ➔ The accuracy of the test dataset is 0.56 and the F1 score is 0.49.

Naïve Bayes:
Naïve Bayes is an algorithm that works on the Bayes theorem of probability which predicts the class of data based on the probability. Naïve Bayes takes data as input and classifies the data based on the Bayes's probability. Naïve Bayes are used to predict the probability of an event to occur based on various factors. For example, If I am buying milk then naïve Bayes can be used to predict that I will mostly buy eggs and bread. We used our patient data which had variables like PrimaryKey, CallConfirmedDT, PatientID, TransportUnitID, TransportLOC, TransportHospitalCode, TransportHospitalName, ReasonForChoosingHospital, PatientAge, PrimaryImpression, SecondaryImpression, PatientDisposition, Distance, Nearest Facility, Distance to Nearest Facility, Success. We split the dataset into train and test, where the training dataset was 70% and the test dataset was 30%. We have a column in the dataset which we calculated based on the distance traveled by each patient after the incident. If this distance is more than the minimum distance, then the Success column has NO which indicates a failure. We applied the NavieBayes classifier provided by R to the train data, which means the model was built using the train data and then We applied the model to the test data and calculated the confusion matrix by comparing the Success column of test data to the predicted data.

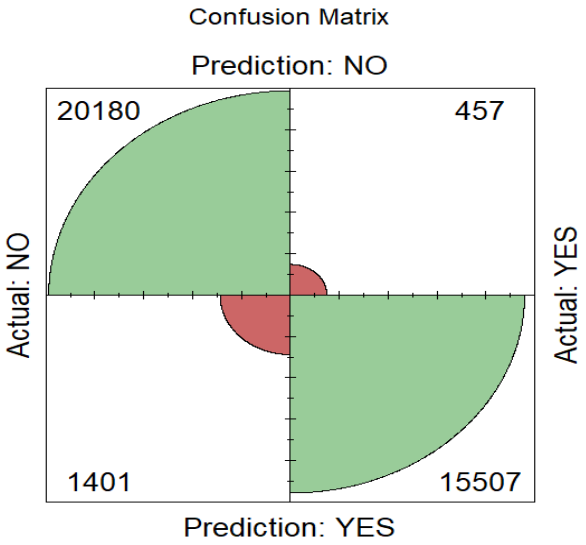


Figure 20 *Confusion matrix visualization for the Naive Bayes model*

Random Forest with Cross-Validation

To estimate the best criteria for the random forest classifier, I chose to use 'GridSearchCV' using multiple inputs for various parameters using {'rf__n_estimators': [100,200,300], 'rf__max_depth': [50,100,500], 'rf__min_samples_split': [2,4]}and also was able to add f1_score metric for model evaluation. We got this feature importance plot when we built the random forest model with Cross-Validation.

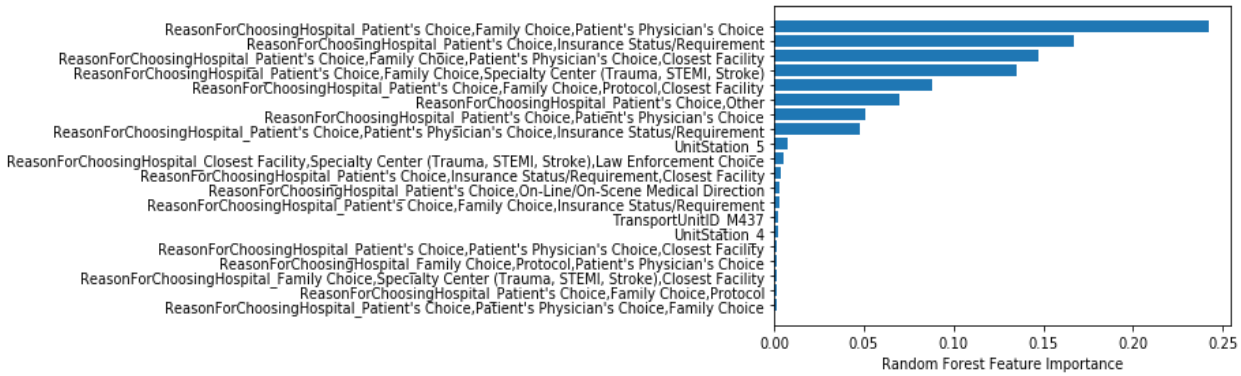


Figure 21 *Feature importance plot for Random Forest model with cross-validation*

This model was built after merging the following data sets

- ➔ Incident
- ➔ Patient
- ➔ Units
- ➔ Facilities

The merged dataset size was (288314, 38) and we cleaned that data by removing the correlated features and removing empty values, and the cleaned dataset had the size of (285024, 347).
To improve the accuracy of the models we have created previously, we used the GridSearchCV function in python which tries to create combinations of train data and calculates the accuracy for each of them to find the combination with high accuracy. So, it basically grid searches the data with cross-validation, it takes input estimator, param_grid, cv, and scoring. This model created 5 split models with 4 different types of parameter combinations, which means the model creates 4*5 = 20 different models.

params	split0_test_score	split1_test_score	split2_test_score	split3_test_score	split4_test_score	mean_test_score	std_test_score	rank_test_score
{'rf__max_depth': 15, 'rf__n_estimators': 25}	0.927210	0.921264	0.916008	0.934924	0.930797	0.926040	0.006729	1
{'rf__max_depth': 15, 'rf__n_estimators': 10}	0.910697	0.910877	0.885758	0.919058	0.925526	0.910383	0.013500	2
{'rf__max_depth': 5, 'rf__n_estimators': 25}	0.752434	0.671365	0.664647	0.699447	0.696694	0.696917	0.030929	3
{'rf__max_depth': 5, 'rf__n_estimators': 10}	0.598002	0.563403	0.534832	0.486954	0.580910	0.552820	0.038998	4

Figure 22 *The top 4 tests scores from the 20 models created in Random Forest using Cross-validation*

This model showed us the best results on the train and test dataset. The model has an accuracy of 0.94 on the training dataset.

	precision	recall	f1-score	support
1	0.80	0.94	0.86	10541
3	0.91	0.73	0.81	9242
11	0.98	1.00	0.99	26614
16	1.00	1.00	1.00	83871
36	1.00	0.98	0.99	228
47	0.76	0.99	0.86	23544
58	1.00	0.05	0.09	393
59	1.00	0.02	0.04	195
95	1.00	1.00	1.00	28924
102	1.00	0.64	0.78	192
107	1.00	0.94	0.97	51
116	0.00	0.00	0.00	4
202	1.00	0.10	0.18	1144
204	1.00	0.08	0.15	212
205	0.80	0.58	0.67	7052
213	0.89	0.33	0.49	4783
215	0.96	0.94	0.95	712
216	0.00	0.00	0.00	10
324	1.00	0.01	0.02	132
325	1.00	1.00	1.00	430
326	1.00	0.10	0.18	124
327	1.00	0.88	0.93	178
328	1.00	0.70	0.82	100
363	1.00	0.38	0.56	13
364	0.00	0.00	0.00	13
374	1.00	0.00	0.01	701
700	0.00	0.00	0.00	3
701	0.00	0.00	0.00	2
999	1.00	1.00	1.00	108
accuracy			0.94	199516
macro avg	0.80	0.50	0.53	199516
weighted avg	0.94	0.94	0.93	199516

Figure 23 *Summary statistics of all the iterations in the random forest model (Train Data)*

The model has an accuracy of 0.93 on the test dataset.

```
1 print(classification_report(y_test,y_test_pred))
```

	precision	recall	f1-score	support
1	0.78	0.91	0.84	4406
3	0.88	0.72	0.79	3986
11	0.98	1.00	0.99	11493
16	1.00	1.00	1.00	35939
36	1.00	0.96	0.98	106
47	0.76	0.99	0.86	10029
58	1.00	0.03	0.05	148
59	1.00	0.01	0.03	78
95	1.00	1.00	1.00	12367
102	1.00	0.47	0.64	89
107	1.00	0.96	0.98	28
116	0.00	0.00	0.00	2
202	0.97	0.06	0.12	485
204	1.00	0.08	0.15	102
205	0.75	0.54	0.63	3075
213	0.85	0.34	0.48	2058
215	0.92	0.93	0.92	312
216	0.00	0.00	0.00	7
324	0.00	0.00	0.00	62
325	1.00	1.00	1.00	173
326	1.00	0.09	0.17	54
327	1.00	0.84	0.91	68
328	1.00	0.67	0.80	36
363	1.00	0.50	0.67	2
364	0.00	0.00	0.00	7
374	1.00	0.00	0.01	332
700	0.00	0.00	0.00	3
701	0.00	0.00	0.00	1
999	1.00	0.97	0.98	60
accuracy			0.93	85508
macro avg	0.75	0.49	0.52	85508
weighted avg	0.94	0.93	0.92	85508

Figure 24 *Summary statistics of all the iterations in the random forest model (Test Data)*

We did some feature engineering on the model and tried to improvise it

- Merged Dataset = Patients + Units + Incidents + Facilities
- Dropped only a few features which were categorical values with large variance, highly correlated features.

- Incident First Due in within or outside Fairfax, Retained the Nan values by replacing with value 1 if the Incident First Due was outside Fairfax County.
- Retaining features such as ‘Unit Station’, ‘Unit Type’, ‘Transport Unit LOC’ as categorical values for analysis.
- Check the model performance and feature importance by hyper-parameter tuning.
- Evaluated the model performance with and without features such as ‘Primary Impression’, ‘Tour of Shift’, ‘Patient Age’.
- The outcome variable used was ‘Transport Hospital Code’

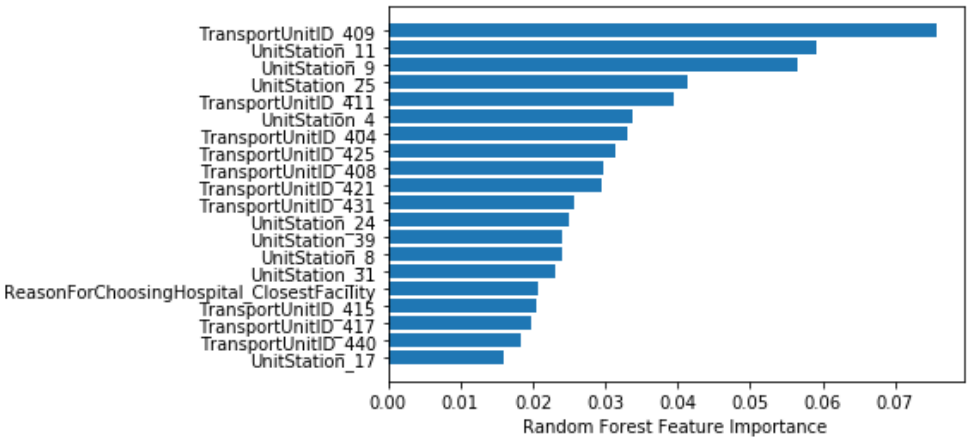


Figure 25 *Feature importance of random forest model*

This was the feature importance plot we got after applying the model, so we used Light GBM boosting framework to improvise our model further which gave us the following feature importance graph

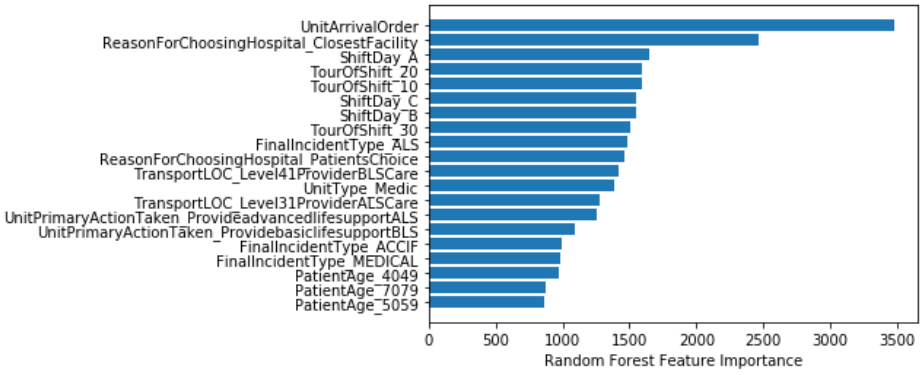


Figure 26 *Feature importance plot after applying LightGBM*

Then the model was analyzed to see how many mismatch percentages there are to find anomalies

➔ Mismatch based on Transport Unit ID

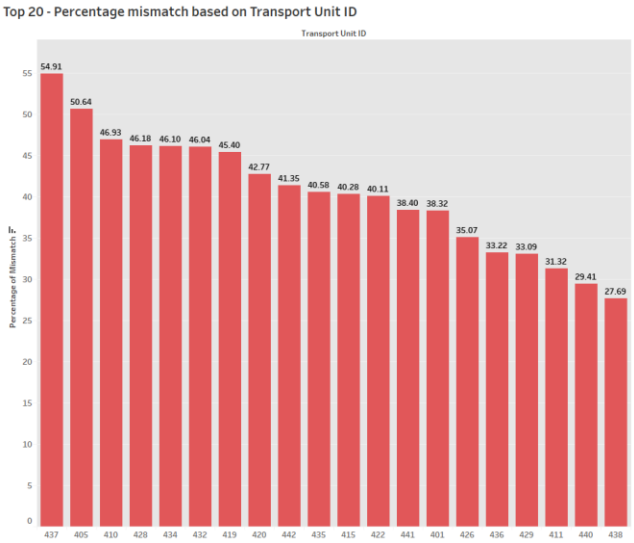


Figure 27 *Visualization describing the mismatch percentage by each Transport ID*

In this graph, we can see the mismatch percentage by each Transport Unit ID, and Unit 437, 405, and 410 have a high percentage of mismatch values.

➔ Match based on Transport Unit ID

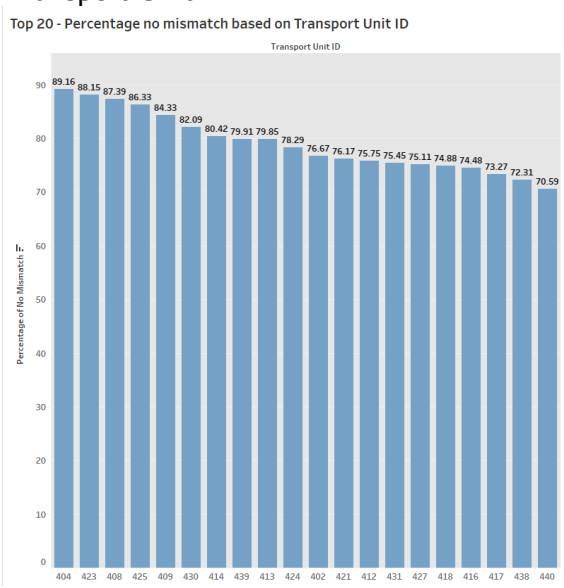


Figure 27 *Visualization displaying the percentage of outcomes from the model that matches the actual output.*

In this graph, we can see the match percentage by each Transport Unit ID, and Unit 404, 423, and 408 have a high percentage of values matched by model.

	PrimaryImpression	total_count	mismatch_count	percentage_of_mismatch
5	Syncope - Syncopal/Fainting Episode (or Near) ...	148	6.0	24.666667
1	Respiratory - Acute Distress/ Breathing Diffic...	311	13.0	23.923077
12	Injury - Head without L.O.C. (S06.0X0A)	109	5.0	21.800000
3	Neuro - Altered Mental Status / Level of Consc...	215	23.0	9.347826
9	Pain - Back (Non-traumatic) (M54.9)	125	17.0	7.352941

And we analyzed the Reason for choosing hospital-wise mismatch percent for 437 units. The primary impression SPeciality Center (TRAUMA, STEMI, Stroke) has a maximum mismatch percent for transport unit 437 which is 13.42%

3.1 Evaluation Metrics:

a. F1 Score: Since our dataset was skewed. We chose to use the F1 score as metric evaluation. The F1 score is the weighted average of the precision and recall, F1 score reaches the maximum value at 1 and minimum value at 0. Both precision and recall contribute equally to the F1 score. Formula F1 score is F1 = 2 * (precision * recall) / (precision + recall)

b. **Accuracy Score:** In classification, the accuracy_score function is used to compute subset accuracy. Whether the predicted labels 'y_pred'match exactly to the corresponding set of true labels i.e., 'y_true'.

3.2 Experimental Results:

Model	Data	F1 score on test	Accuracy on test
Model 1 - RF	Patients, Units, Incidents, facilities datasets merged with distance between each Unit ID and facility	0.70	74%
Model 2-RF	Patients, Units, Incidents and facilities datasets merged	0.91	93%
Model 3 - DT	Patients, Units, Incidents and facilities datasets merged	0.61	67%
Model 5 - DT	Patients, Units, Incidents, facilities datasets merged - Only Top 20 important features	0.62	67%

Model 6- RF	Patients, Units, Incidents, facilities datasets merged - Only Top 20 important features	0.93	94%
Model 7- RF	Patients, Units, Incidents, facilities datasets merged - By removing top 3 influential features	0.64	71%
Model 8- LightGBM	Patients, Units, Incidents, facilities datasets merged - By removing top 3 influential features	0.7	69%
Model 9 - LR	With 2 features - TransportUnitID, TransportLOC	0.7	72%

4. Visualizations

- Shift Day/Tour of Shift

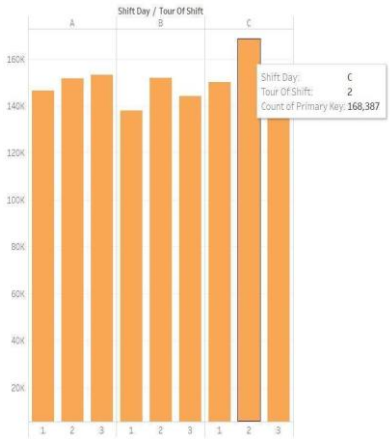


***Figure 28**: Shift Day/Tour of Shift*

The units in FRD work in 9 shift days in 3 tours. The schedule is as follows.

SUN	MON	TUE	WED	THU	FRI	SAT
A day 1	C day 3	A day 2	B day 1	A day 3	B day 2	C day 1
B day 3	C day 2	A day 1	C day 3	A day 2	B day 1	A day 3

***Figure 29:** Shift Schedule (Department, 2021)*

In the figure 28,29, we can see that units work in 3 shifts which are A, B, C, and each shift will have 3 tours over a period of 9 days. The visualization shows us which shift, and tour of the shift had the most incidents, from the data we have we can see that shift day C on its second tour has experienced the most incidents over a period of 3 years which is around 168,300 incidents, but these numbers don't vary much so every day and tour of shift has handled approximately same number of cases.

- Transport Level of Care

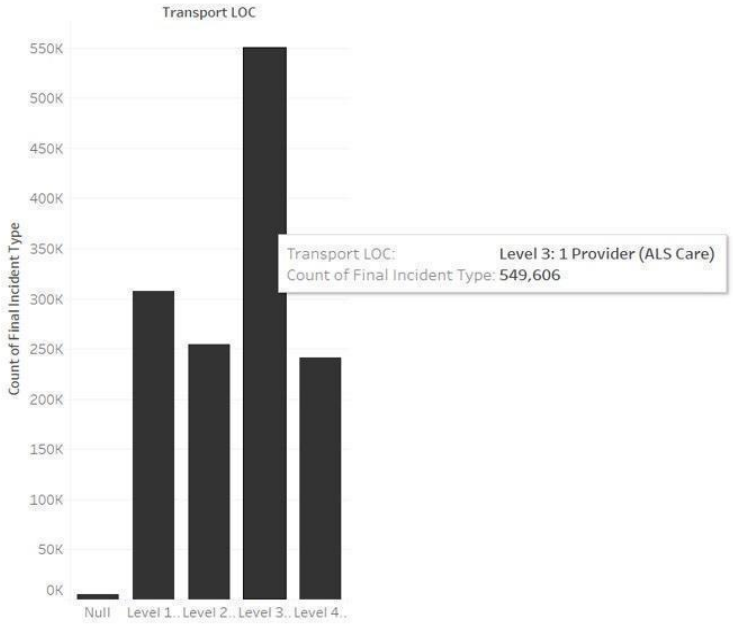


***Figure 30**: Level of transport visualization*

The level of care offered during transport to the patient is divided into 4 levels, it is categorized based on the number of medics present with the patient. This is decided on the type of incident that occurred.

Level	Number of medics	Examples of incidents
Level 1	Three or more	Cardiac or respiratory arrests, combative head injuries, or multi-system trauma patients.
Level 2	Two	Acute STEMI, CPAP, or seizure patients with potential for repeat seizures.
Level 3	One ALS	Resolved hypoglycemia, non-bloody vomiting with normal vital signs, abdominal pain without tenderness and with normal vital signs.
Level 4	One BLS	MVC neck-back pain, general injuries or illness patients.

***Figure 31**: Level of Transport Care (Department F. C., 2019)*

As we can observe in the visualization, the level 3 care which includes one ALS medic was the most needed level of care in 3 years, 549606 incidents needed level 3 medic care. ALS medic

means nothing but Advance level of care. From the above graph we can conclude that in the maximum number of accidents Level 3 care which includes one ALS medic is provided.

- Unit Types

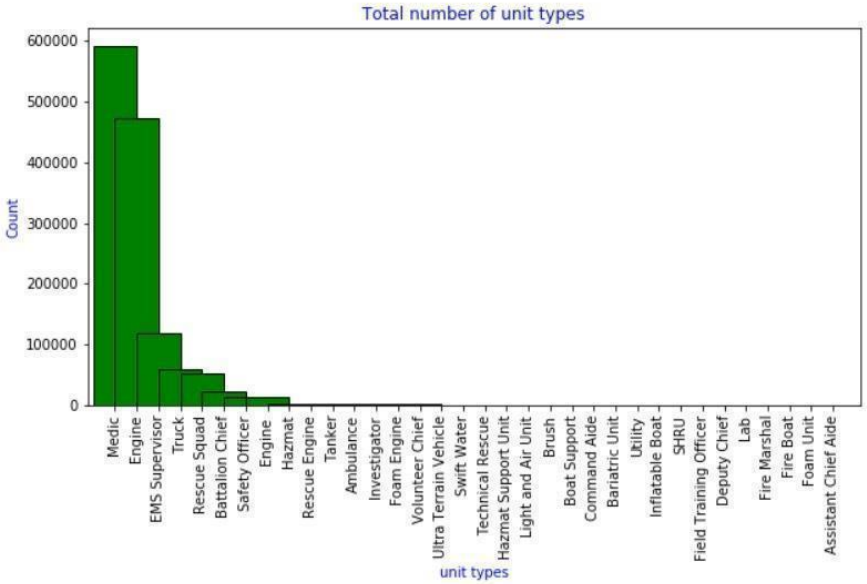


***Figure 32**: Unit types of vs count of each one called in incidents*

The unit type in FRD dataset represents the type of unit which was called when an incident occurred. There are many types of units mentioned in the dataset each serving different purposes. We can clearly see in the visualization (Fig 32) that the medic unit was called for the maximum number of times, the number goes up to 600000 times. This means approximately 40% of the time the medic unit is called whenever an incident occurs. The top 5 most called units are Medic, Engine, EMS Supervisor (Emergency Medical Services), Truck, and Rescue Squad.

- Correlation Plot

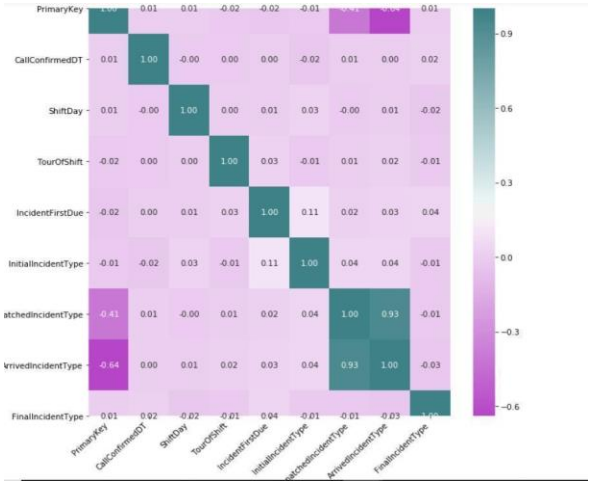


***Figure 33**: Correlation Matrix.*

The correlation matrix (Fig 33) is used to find relations between the features of a dataset. We don't see any correlation which would make sense in the visualization. The dispatched incident type and arrived incident type show a correlation that does not have any higher importance.

- Average Distance vs Minimum Distance

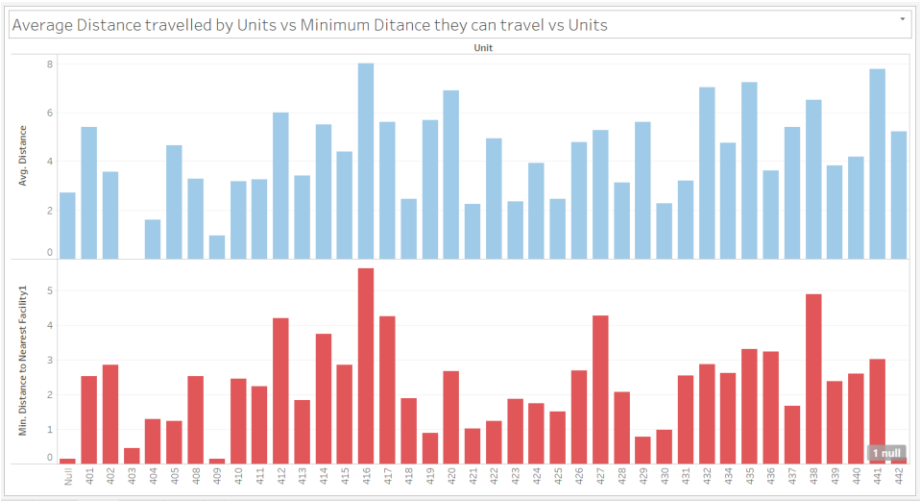


***Figure 34**: Average Distance vs Minimum Distance by units*

This graph (Fig 34) basically shows us if any unit is traveling more than needed because the graph on the bottom shows the minimum distance one unit can travel, which also means that it shows us the distance between the unit station and the nearest facility. And the graph on top shows the average distance unit travelled by every unit.

- Unit wise count of Success

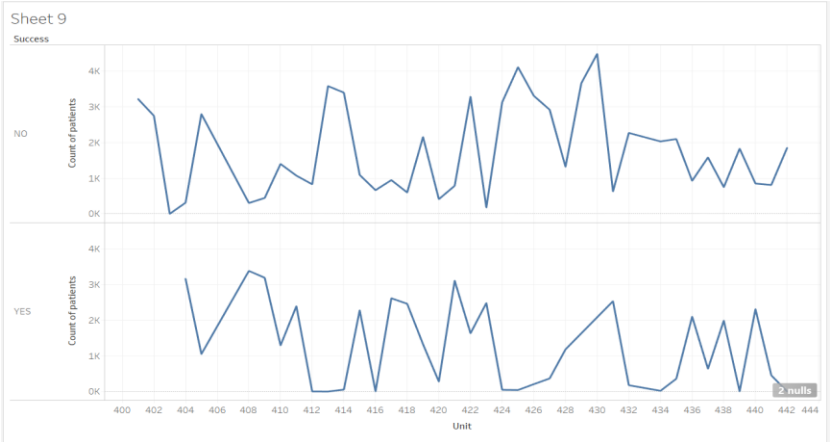


***Figure 35**: Unit-wise count of patients in YES and NO success.*

In this graph (Fig 35), we have tried to show the count of patients handled by each unit in both scenarios YES and NO. This means the count of patients which were taken / not taken to the nearest hospital by each unit. This is the column which is calculated based on how much distance units travelled to take patients to hospitals and the average of distances was compared with the

minimum distance the unit could have travelled to reach the nearest hospital. And then it's decided if the case was successful or not. The column has YES or NO values. YES, means the travelling situation of the unit was success and NO means it was not.

- Units' vs Counts of incidents vs avg distance travelled

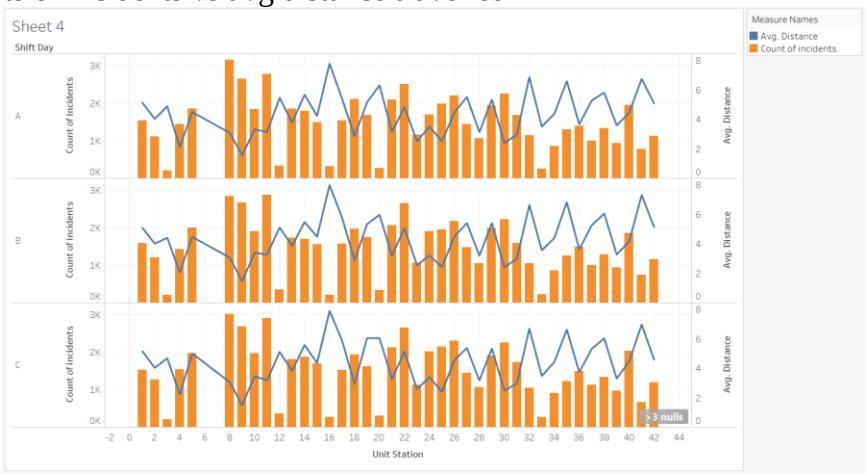


***Figure 36**: Unit stations vs Count of incidents vs average distance traveled*

In this graph (Fig 36) we can see that, when any unit station has a higher number of handled incidents it has traveled a lower distance on average and vice versa.

- Transport Level of Care in every year

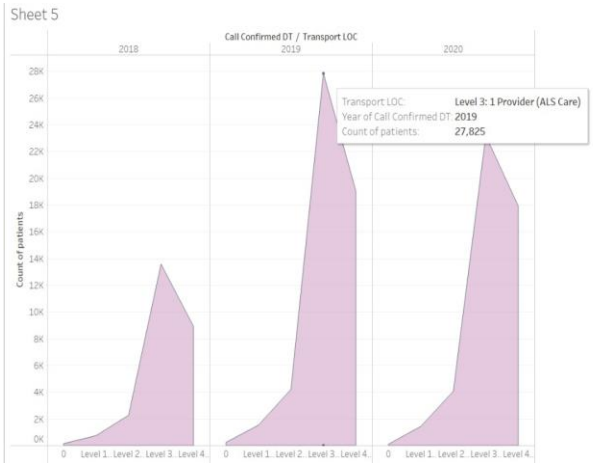


***Figure 37**:* Transport Level of Care in the years 2018, 2019, and 2020.

The above graph (Fig 37) depicts information about the Transport Level of Care in the years 2018, 2019, and 2020. The Y-axis gives information about the count of patients corresponding to the various levels of care in each of the above-mentioned years. In the year 2018, the highest cases of victims were given ALS care which is of level 3 Level of care and the lowest cases that year belong to level 1 Level of care. In the year 2019, the highest cases of victims were given ALS care which amounted to about twenty-five thousand cases which are corresponding to the level 3 Level of care and the lowest cases that year belong to the level 1 Level of care accounting for two thousand

cases. There has been a comparative rise in the number of cases reported over the years. In the following year 2020, there has been a decrease in the number of cases reported when compared to the year 2019. Following the pattern as in the years 2018 and 2019 the year 2020 also has the highest number of cases in ALS care (Level 3). To summarize this plot, we can observe that the highest number of cases reported was in the year 2019 and the Level of care given to the victims was ALS care. Irrespective of any year or circumstance the most common approach to provide care to a victim would be ALS care.

- Mismatch Percentage and Count for every Hospitals

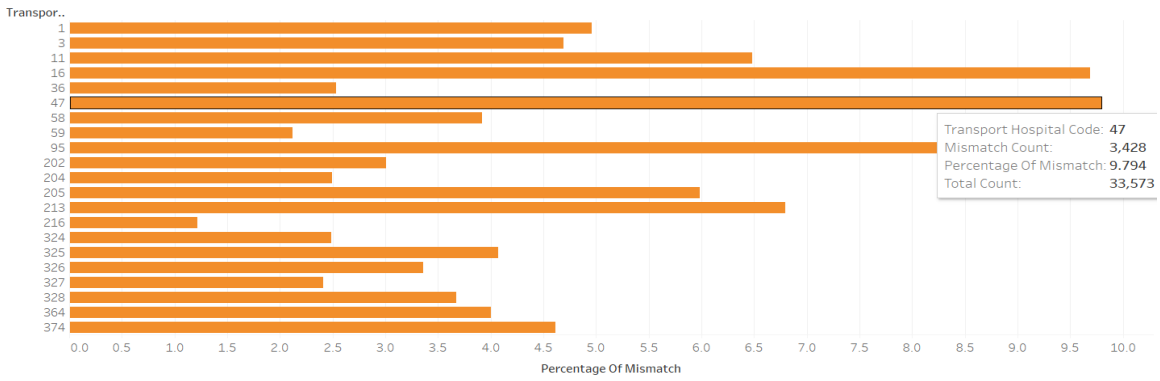


***Figure 38**: Mismatch percentages and counts*

The Percentage of Mismatch is on the X-axis, and the Transport Hospital code is on the Y-axis, as shown in the graph (Fig 38) above. The graph indicates that the Transport Hospital code 47, has the largest proportion mismatch which is 9.79 percent. Furthermore, we can observe that the chart demonstrates that the Transport Hospital code 47, has a mismatch count of 3,428 and a total count of 33,573. Whereas Transport Hospital code 216 has the lowest percent mismatch.

- Mismatch Percentage and Count for every Units

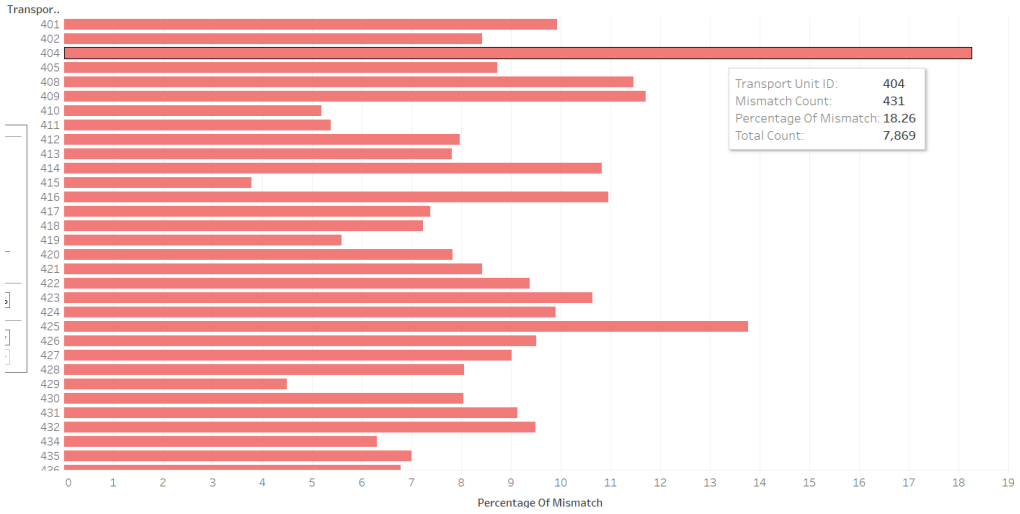


***Figure 39**: Mismatch percentages and counts*

As indicated in the plot (Fig 39) above, the Percentage of Mismatch is on the X-axis, while the Transport Unit ID is on the Y-axis. According to the histogram, Transport Unit ID 404, does have the highest percentage discrepancy of 18.26%. Additionally, we could see that the Transport Unit ID 404, has a mismatch count of 431 as well as a total count of 7,869 in the chart. The Percentage mismatch for the Transport Unit ID 415 has the least Percentage.

- Mismatch Percentage and Count for Transport Level of Care

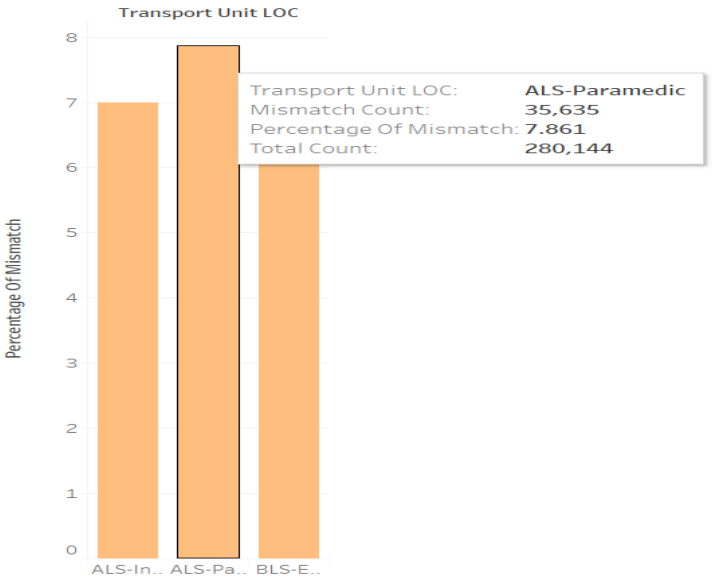


***Figure 40**: Mismatch percentages and counts*

The above graph (Fig 40) depicts information about the Transport Unit Level of Care. The Y-axis gives information about the percentage of mismatch corresponding to the various levels of care and the X-axis gives information about the Transport Unit Level of Care. Where the highest percentage of mismatch was seen in ALS- Paramedic Transport Unit Level of Care with 7.86 percentage.

- Mismatch Percentage and Count for every Unit Primary Action taken

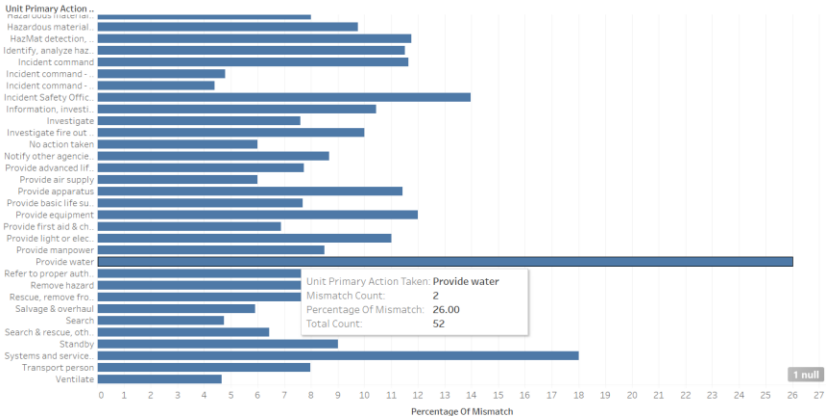


***Figure 41**: Mismatch percentages and counts*

From the illustration (Fig 41), we can see that the Percentage of Mismatch is on the X-axis and the Unit Primary Action is on Y-axis. Where the graph depicts that the highest percentage mismatch in the Unit Primary action Taken is for the Provide water action that is 26%. Moreover, we can also see that the graph provides the mismatch count, which is 2, and the total count which is 52 for the Unit Primary Action Provide water. The second-highest percentage of Mismatch is with the Unit Primary Action, Systems, and service with 18 percent. And the least percentage mismatch is for the Unit Primary Action named Incident command.

- Facilities frequently visited by Units

***Figure 42**: Dashboard 1*

The dashboard (Fig 42) shows us the frequently visited facilities by units. Whenever any incident occurs fire personnel take the victims to hospitals. This visualization focuses on the top three hospitals preferred by every unit. In the above screenshot we can see that unit number 417 takes the patient to two hospitals 11, 16, 95 and unit 417 has taken patients to hospital 11 total 2612 times, as hospital 11 is closest to unit 417 the finding makes sense. We can see the same visualization for every unit. In the above dashboard we also have three bar graphs in the bottom. The first paragraph indicates the count of match and mismatch predicted by our model. According to the bar graph for unit 417 there was a mismatch for 487 records. The second bar graph in red shows us the number of times a unit preferred a particular hospital so in the case of unit 417 hospital 11 was preferred 2612 times. and the last bar graph can be used to see the distance between units and those hospitals. This dashboard is very helpful because it can help our partners find anomalies in units. For example, if a particular unit is selected and we see that the unit is traveling to a hospital which is far away from its location for a maximum number of times then this can be considered as an anomaly. We must understand that inclined behavior can sometimes be justified,

for example if a unit took patients to a particular hospital a small number of times it is possible that the patient had some type of injury which can only be cured by a particular hospital in this case the situation won't be considered as an anomaly. But as we have less knowledge about specialties of hospitals and the necessities of the situation, we can't be the judge of the situation, so handing this dashboard over to partners was the conclusion because partners can further filter out units and analyze what is correct behavior and what is wrong behavior.

- % Mismatch based on Transport Unit ID and related features

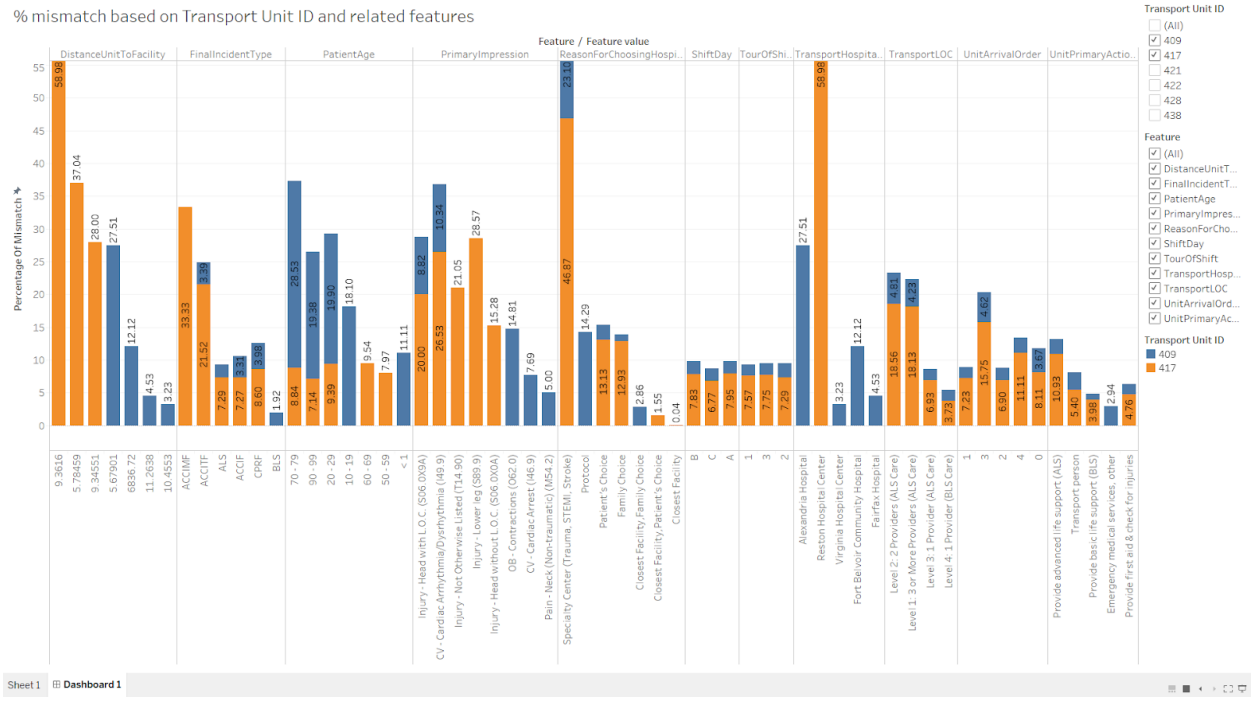


***Figure 43**: Dashboard 2*

The model which was built by our team predicted the transport hospital name by taking all the data as input. There was a certain percentage of mismatches in prediction. This data which showed mismatch had various influential features and the dashboard shown above is used to indicate the same. There were specific unit ID's (fire station units) who's behavior was inclined from normal behavior and our partners wanted to have a closer look at these unit ID's. So the purpose of building this dashboard was to have a closer look into those unit IDs and the mismatch percentages of each feature. So, in the dashboard we can select unit ID and the features we are interested to look into, percentage mismatch of the combination is displayed. If in the future or unit ID are added into the dashboard then we can see the big picture by combining both visualizations. To explain in detail in the first visualization if we find that particular unit ID is preferring to go to a particular hospital then we can filter the same unit ID in the second visualization to see what features are mismatched and are influencing the transport decision-making.

- Transport Unit IDs with the number of incidents they handled

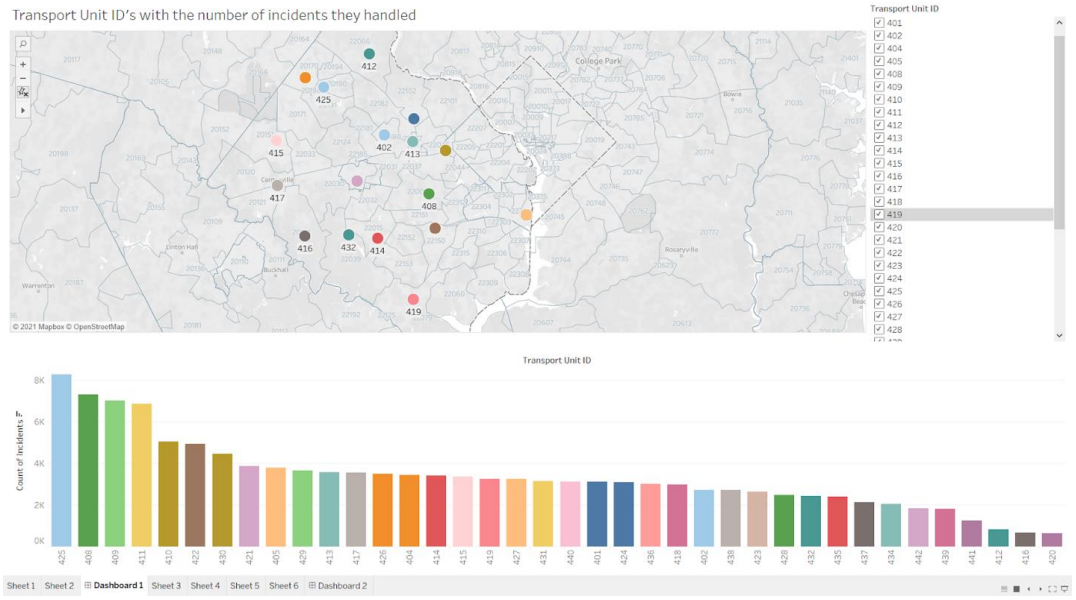


***Figure 44**: Dashboard 3*

The dashboard (Fig 44) above shows the geographical locations of all the Transport Units in Fairfax County. Each transport unit is represented in each color. According to the number of cases handled by each of the transport units, their sizes can be defined. For example, the Transport Unit 425 is represented in light blue color and has the highest count of cases handled, that is, around 8K. Similarly, Transport Unit 408 is represented in green color and has the second-highest count of cases handled.

- Minimum Distance vs Average Distance

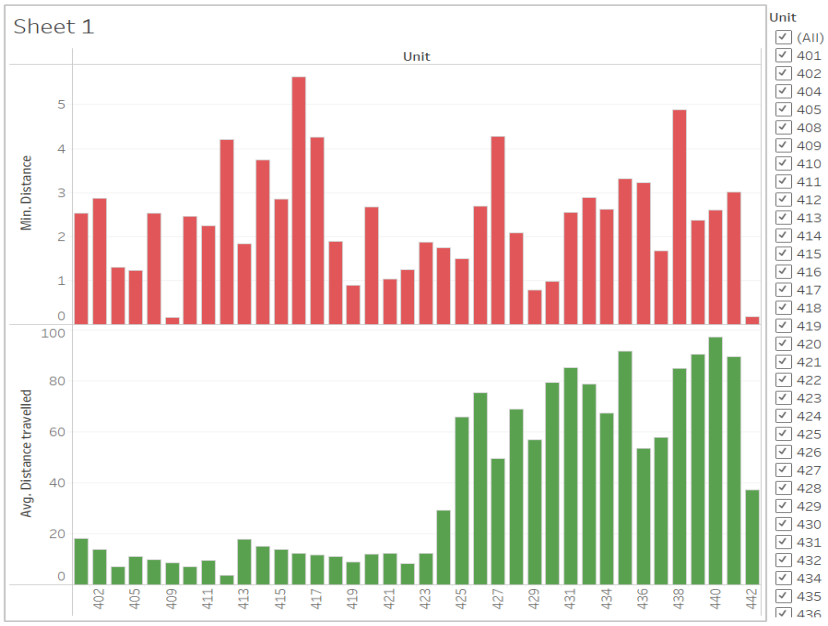


***Figure 45**: Dashboard 4*

The interactive dashboard (Fig 45) above determines the distance between the transport unit to the nearest hospital. By using the lat long data the minimum distance between each unit and the nearest facility was determined. From the model, we were able to determine the distance of each unit for each case, and their average was being defined. Later compared the average distance and minimum distance. From this data, we can compare the average distance to the minimum distance they could have traveled to reach the facility.

5. Findings

Initial analysis was done on both the individual tables and facilities datasets. The main question of our project is to find out the features which are important and play a role in the decision of where to take a patient when any incident takes place.

Finding out the best hospital to go to when any incident takes place, is one of the major findings that helped us to proceed in the project. This helped us determine what records in the dataset can be classified as successful and what records can be classified where there is a scope of improvisation. For example, we must see the location data of an incident and then check it with the facilities data to understand which is the nearest hospital. The second thing they must consider is the type of incident, for example, if there is a burn involved in the incident, the Fire and Rescue Department have to take the patient to a hospital having the facility to treat burns, if the patient is facing trauma, they have to take the patient to a trauma center.

The nearest hospital from the location of the incident is another major finding. For privacy reasons, the fire and rescue department did not provide us with any incident locations. We were suggested to use the location of the fire station which offered help to the victims. This makes sense because whenever any incident takes place, the fire station nearest to that location is alarmed and asked to help. This data about locations of all the fire stations was stored in the form of latitude and longitude and this can be referenced with the location data of the hospitals of Fairfax to find the nearest hospital.

How is the rate of case measure? This helped us identify categories such as stations, injury type, treatment, etc where certain groups are performing better or worse than others. The features which we used are: ShiftDay, TourOfShift, InitialIncidentType, UnitPrimaryActionTaken, TransportHospitalCode, ReasonForChoosingHospital, PatientAge(bucketed), MedicationGiven, Facility_Location_Code, Facility_Type_Of_Facility. The logic that the fire department is following is Emergency Medical Dispatch cards where there are steps and procedures that unit personnel must follow once an incident has been reported.

The zones/areas in Fairfax that experienced a major number of incidents. We visualized the hot zones where we see a lot of activity on the map. We used the Hospitals as pins and then colorized the zip codes based on the count of incidents in a zip code, eg: the zipcode where there are a lot of incidents can be red and the zipcode with fewer incidents can be blue/green. From the Hospital's datasets, we used the Facility_Full_Address field to extract the zip code of the hospital. Eg: 15225 Heathcote Blvd, Haymarket, Virginia, 20169, United States. We splitted the address by commas and get the zip code as the second last field. As for where the incident took place, we used the Hospital where the patient was taken to identify the area. We used the column TransportHospitalCode from the Incident dataset to connect to the Facility_Location_Code in the Facilities data set. Further, we can take into consideration the population density of a zip code to normalize the data. A larger population will mean more incidents, so we can divide the count of

the incident in a zip code by its population to understand which areas have a high incident rate. This study will help us identify the areas where we need to focus more and maybe set up more hospitals or response units to improve the response.
The problem statement of the project majorly focuses on transporting the patient to the best facility as discussed earlier. By working with the visualizations, we have found out some discrepancies and anomalies in the data.

Unit 409

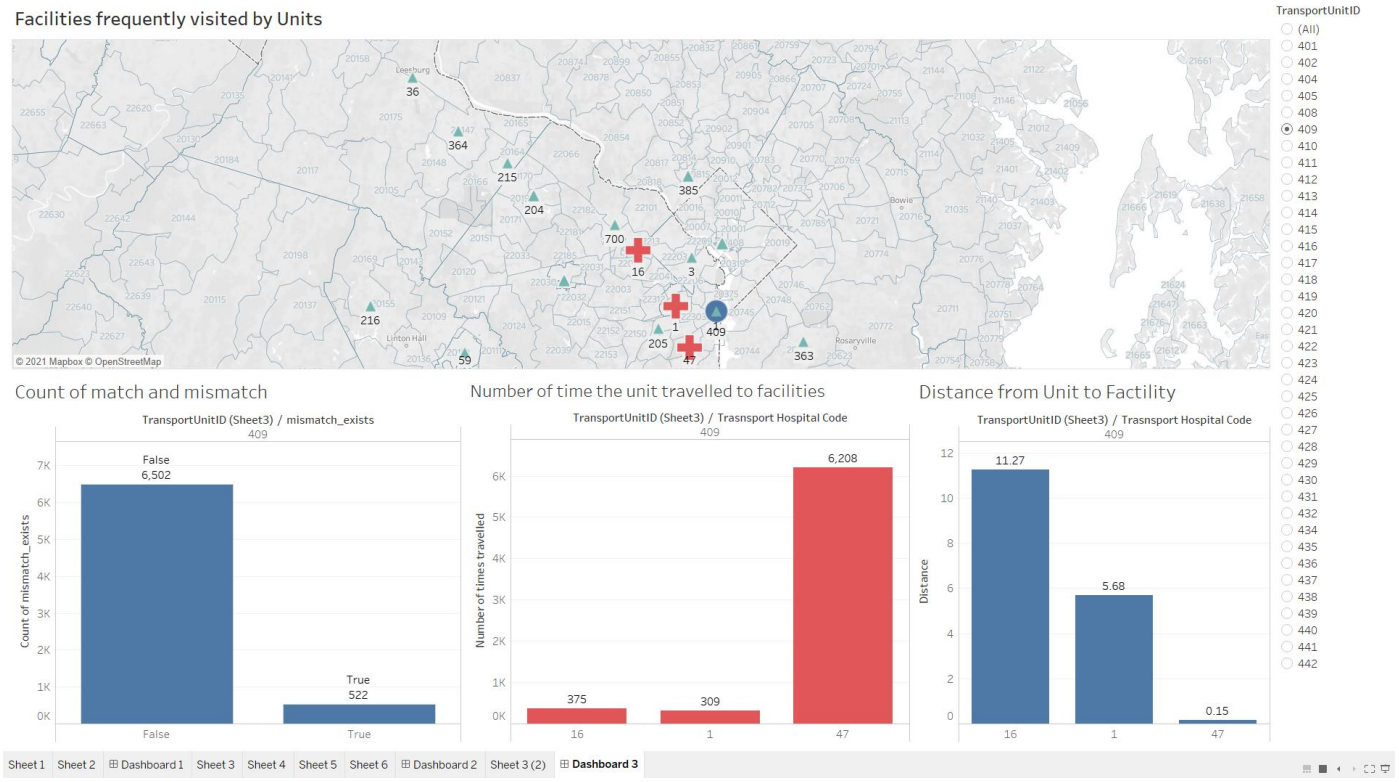


***Figure 46**: Unit 409*

If we closely observe this visualization, we can observe following things

1. Unit 409 has visited hospitals 1, 16 and 47 very frequently.
2. Hospital 3 and 205 were much closer to unit 409 than hospital 16.

Which means that there are some anomalies when it comes to unit 409 and unit 409 is inclined towards some hospitals

Unit 417

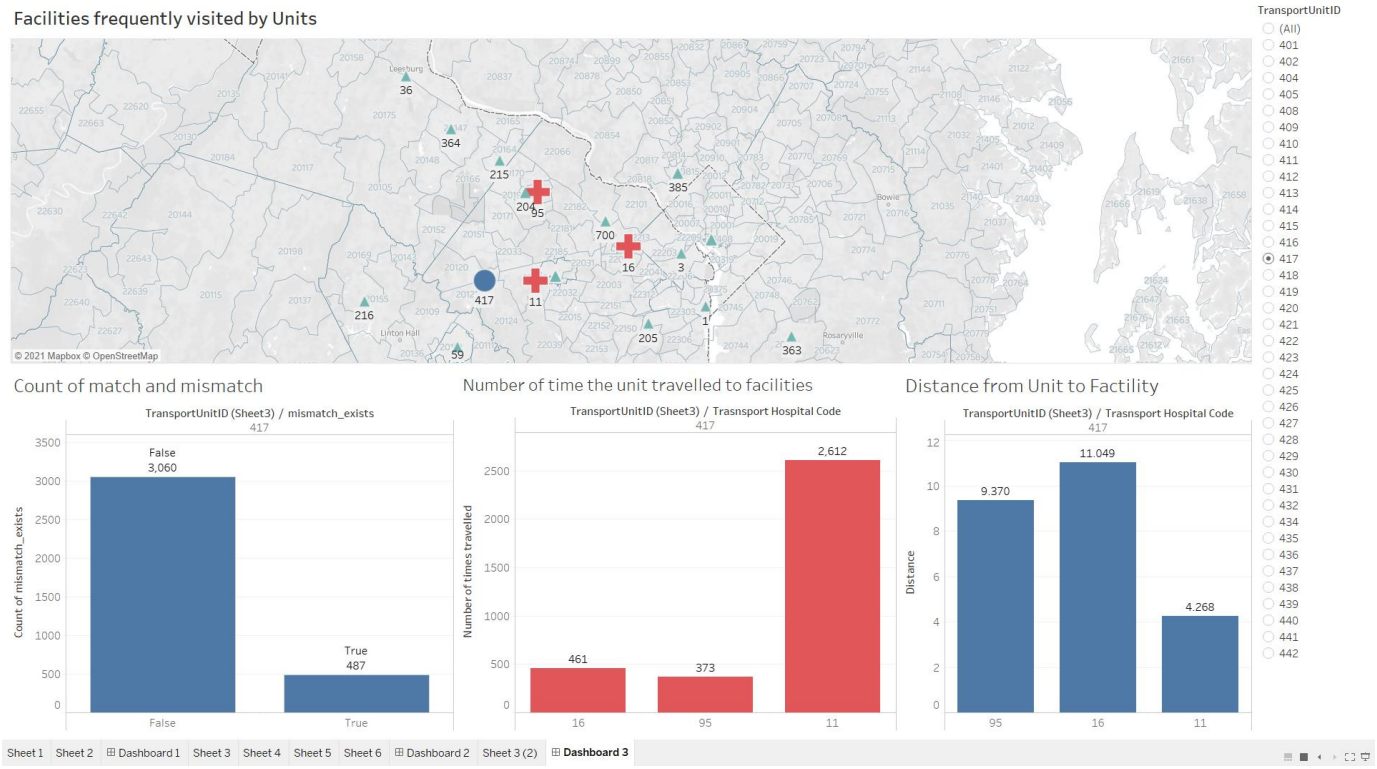


***Figure 47**: Unit 417*

If we closely observe this visualization, we can observe following things

1. Unit 409 has visited hospitals 11, 16 and 95 very frequently.
2. Hospital 59 was much closer to unit 417 than hospital 95.

Which means that there are some anomalies in the case of unit 417 too. To closely look at these anomalies and the features which can have potential to influence the transport decision making we can have a closer look at the second visualization '% mismatch based on Transport Unit ID and related features'

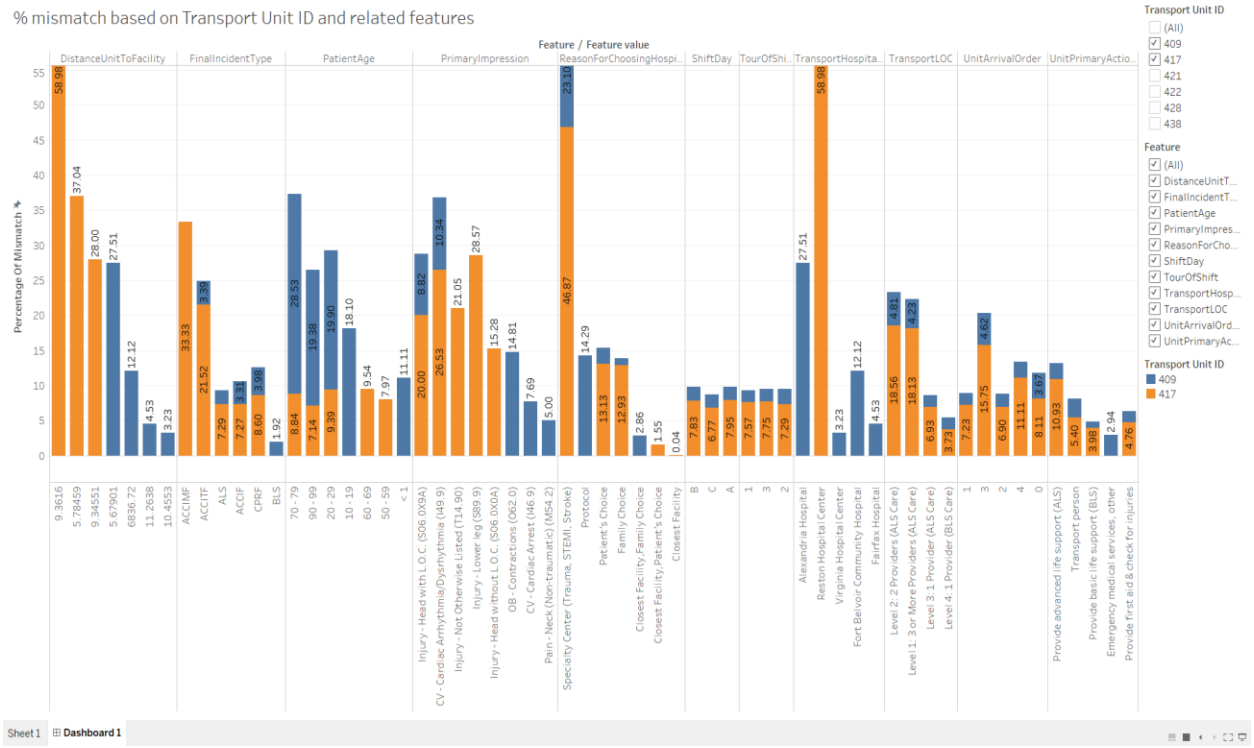


Figure 48

Following findings can be observed from this visualization which shows percentage mismatch for Unit 409 and 417 for every feature and some attributes in the features.

1. Patient Aged 70 - 79 influences the decision for both units by a very high mismatch percentage. For unit 409 the decision is influenced by 28.53% and for unit 417 it is by 8.04%
2. Reason for choosing hospital - specialty center - also influences both the units' transport decision for unit 409 the decision is affected by 23.3% by reason for choosing hospital - specialty center and for unit 417 its affected by 46.87%
3. Transport Hospital - Alexandria Hospital - influences the decision by 27.51% for unit 409
4. Transport Hospital - Reston Hospital Center - influences the decision by 58.98% for unit 417

6. Summary

Initially, the problem at hand was to help improve the transport decision-making process in the Fairfax County Fire and Rescue department. For this, we were provided with the data of all the incidents that took place over the last 3 years. The data has details of all the units, services, facilities, and incidents. Using this data, and applying some methods to it, we found the most influential attributes that affect transport decision-making we're DistancefromUnitToFacility, UnitStation, TransportUnitID, IncidentFirstDue, ReasonForChosingHospital, ShiftDay. We later used these attributes in the machine learning models that we built. Eventually, after many iterations, we built a model with 92% accuracy and also dashboards that provided the patterns and trends in the data and thereby giving the firemen suggestions of which facility to choose based on the various incident types and the location of the incident.

Accuracy is high but that does not mean that the model is good. Since the data is skewed the minority class is wrongly predicted. This is captured correctly in the F1 score. By adding distance features to the unit dataset, the model was able to perform better. Random Forest outperformed the LightGBM, Logistic Regression and Decision Trees. We drew conclusions from model predictions to find anomalies and further evaluate how various attributes contributed to these decisions. The model probability (confidence threshold) and mismatch percentage were considered to narrow down the anomalies from the dataset. We found few of the units tend to travel more than the average distance. There was a higher likelihood of visiting certain hospitals more compared to the hospitals nearby. We summarized it using the tableau visualization. Whenever the reason for choosing a hospital was a patient's choice and specialty center was trauma stem, stroke the percentage of mismatch for such categories increased.

7. Future Work

The future and undiscovered scope of this project can be following things

1. Adding data to visualization: Adding all the unit data in '% mismatch based on Transport Unit ID and related features' visualization. This will give a better idea of anomalies in the data for every feature.
2. Add more features: Add the medications and procedures data sheets into the model predictions and check how the medications and procedures influence the transport decision making.
3. Ensemble model: Build a lot of models in distributed data and combine them using LightGBM or XGBoost.

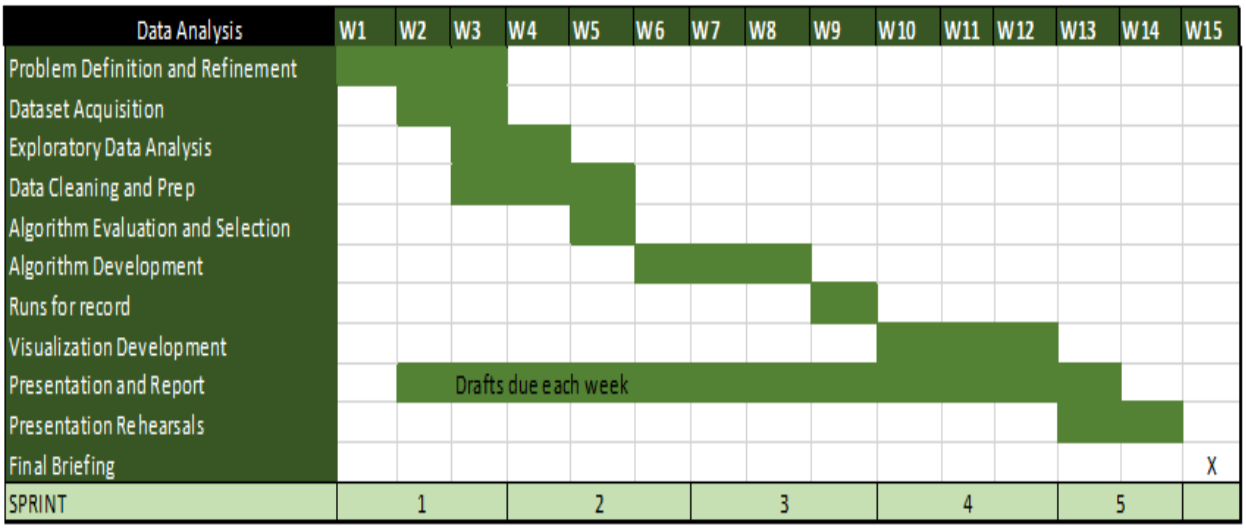
8. Appendix A: Code References

Link for Project code: Github

9. Appendix B: Risk Section

Week	Risk Name	Description	Impact	Actions
1	Lack of patients' data	Patient age and gender data is not present in the data	High	We asked partners for that data
7	Data Preprocessing	As we were handed new data, we still have a lot to do in terms of pre-processing	High	We had winter break to catch up with the schedule
7	Cross Referencing data	Incidents, patients, facilities data must be merged to find the distance value, we might run into problems there	High	The problems occurred during merging data were not that huge and were handled.
9	Improvising models	We are trying to add maximum features to our model while maintaining its predicting accuracy	High	The models were created by adding features one by one creating variations
11	Improving Visualization	Where are trying to create interactive and reusable visualizations to provide it to partners	High	The reusable and interactive visualizations were created and delivered to partners

10. Appendix C: Agile Development



11. References

[1] Department, D. a. (2021). *George Mason University DAEN 690 Data Dictionary Fall 2021, Second ALS Provider.*
[2] Department, F. C. (2019). *EMS Manual.*
[3] “Conda," Vers 4.9.2. Anaconda Software Distribution, 10 November 2020. [Online]. Available:
https://docs.anaconda.com/.
[4] “Python," Vers. 3.8.8. Python Software Foundation, 24 September 2020. [Online]. Available: https://www.python.org.
[5] P. Krishnamurthy, "Towards Data Science: Understanding Data Bias," 11 September 2019. [Online].
Available: https://towardsdatascience.com/survey-d4f168791e57. [Accessed 8 March 2021].
[6] W. Goodrum, "Elder Research: Statistical & Cognitive Biases in Data Science: What is Bias?" 6 October 2017. [Online]. Available: https://www.elderresearch.com/blog/statistical-cognitive-biasesin-data-science-what-is-bias/. [Accessed 8 March 2021].