A picture containing icon

Description automatically generated**A picture containing logo

Description automatically generated**

**Term Project**

**Data Management of Opioid toxicity-related deaths and hospitalizations**

**Date:** April 3, 2023

**Course Number:** SCS 3252

**Section Number:** 051

**Course Name:** [Big Data Management Systems & Tools](https://q.utoronto.ca/courses/301421)

**Instructor:** Stephen Giles

**Submitter:** Irina Belaya

1. **Executive summary:**

Canada has severely experienced the “opioid-related” deaths and other harms with tragic impact on recipients, their families, society, and ultimately the health care system (1). These impacts are results of multiple factors. Evidence demonstrates the impact of covid-19 pandemic on aggravating “opioid-related” overdose, death, and other harms. Overall, opioid related Emergency Medical Services (EMS) was higher than deaths in Canada from 2018 to 2020. AB, MB, followed by Ontario led the order. Ontario, however, showed slightly (insignificantly) higher deaths. The data shows that the highest number of deaths are among the young to middle age group (20 to 59 years old). my analysis also shows that Ontario opioid related Emergency Medical Services and deaths are equal among females and males.

Place of residence, gender and age group are strong predictors of opioid/stimulant toxicity outcome. Random Forest algorithm is recommended as the most effective machine learning model for further analysis of opioid/stimulant toxicity outcomes.

Based on my analysis it is imperative to implement more vigorous regulations and oversee prescription. my analysis also indicates that attention of the underlying causes is required, and potentially developing and implementing social support programs for those groups at higher risk of opioid/stimulant toxicity episodes would be beneficial.

1. **Introduction:**

Canada’s opioid crisis is complex and multifaceted. The current overdose emergency, driven primarily by a rapid increase in the use of fentanyl, a powerful synthetic opioid that is similar to morphine but is 50 to 100 times more potent (2) and other powerful illegal opioid drugs, has led to an unprecedented number of overdose deaths; but this crisis reaches far beyond the illegal drug market. For many Canadians, this crisis has its roots in high levels of addiction to legal opioids, caused in part by inappropriate prescribing practices and poor education about the risks associated with opioids. For others, substance use disorders have much deeper roots in trauma, social and economic inequities, and mental health issues (1).

1. **Objective:**
2. To identify the overall trends in opioid toxicity-related hospitalizations and death
3. To find whether there is any correlation among demographic characterizations with hospitalizations and deaths.
4. To detect the major factor(s) contributing to opioid toxicity-related deaths in Canada
5. **Materials and methods:**

We used federal government of Canada data repository available at

<https://health-infobase.canada.ca/substance-related-harms/opioids-stimulants/technical-notes>.

We engineered the database using preliminary explorations, visualization, and cleaning using Databricks community edition. Finally, I have tested few machine learning models to find the best fit for the project objectives.

* 1. **Data Engineering:**
     1. **Data description and early exploration**

The database is composed of 11 columns and 14938 rows. The values are as follow (3).

**Substance:** 4 categories of Fentanyl & fentanyl analogues, non-fentanyl opioids, Stimulants and other psychoactive

**Source:** deaths, hospitalizations, emergency medical services (EMS)

**Specific\_Measure:** A data breakdown which encompasses age, sex, type of opioid, involving stimulants or other psychoactive substance, age group or type of opioid by sex

**Type\_Event:** Total apparent opioid toxicity deaths, accidental apparent opioid toxicity deaths, and suicide apparent opioid toxicity deaths

**Region:** Canada, provinces, territories

**Pruid:** period ID

**Time\_Period:** 2016-2021

**Year\_Quarter:** year quarterly

**Aggregator:** It contains year, age group, type of opioids, gender (male or female)

**Disaggregator**: It contains year, age group, type of opioids, gender (male or female)

**Unit:** It shows Crude rate, number, and age-adjusted rate

**Value:** It is the numbers representing the count of events (death or EMS)

* + 1. **Data cleaning:**

The initial data investigation revealed the following issues:

1. The dataset included counts, percentages, crude rates, and age-adjusted values of death, hospitalizations, and Emergency Medical Services (EMS). Data reported in ratios, crude rates, and age-adjusted values needed more information to disaggregate; therefore, it should be omitted from the final dataset. Numbers will be used as a unit of measure.
2. Provinces and territories address the confidentiality of sensitive information by suppressing small counts. Those records were reported as negative numbers and should be omitted from the final dataset.
3. Manitoba and Territories report their data for several distinct regions. Those regions should be combined under their respective province or territory for the final dataset.
4. Data reported for Canada should be removed from the final dataset to avoid double counting.
5. The analysis will include yearly data; hence the quarterly records should be omitted from the dataset.
6. Data provided for 2017 and 2022 was incomplete and should be removed from the dataset.
7. The dataset included total records for several types of events (death, hospitalization, emergency services), and those records should be omitted from the dataset to avoid double-counting.

The indicated issues were addressed during the data-cleaning process. The data cleaning process included the following steps:

1. First, I uploaded the dataset csv file into Databricks, and created a dataframe called “df” which can be accessed through the following link:

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/8044888775324931/3765691144116402/5854611345732810/latest.html>

1. Next, I changed the column “Unit” to number showing as “df2” and then selected values higher than -1 and named the dataframe as “df3”.
2. When df3 is grouped by region, I can see in Fig.1 there some provinces are divided in multiple regions. For example, there are regions Manitoba and Winnipeg Manitoba as well as Territories, Yellowknife Northwest Territories and Whitehorse Yukon. The sub-divided regions were categorized by provinces and the territories were grouped as territories. A column ‘new\_region’ was created where all related cities were grouped under their respected provinces and territories. As result I only had 10 unique provinces and one territory category. This ensures that there are no overlaps or double/partial names (Fig. 2).

Table

Description automatically generated

**Figure 1.** The outcome of GroupBy function in all regions.

Table

Description automatically generated

**Figure 2.** New-region column with 10 provinces and 1 territory

4) Next, I decided to exclude data from 2017, and 2022 due to the fact that it covers the first half of the year so it is not complete.

5) The summary rows containing aggregate data, such as 'Total apparent opioid toxicity deaths', 'Total opioid-related poisoning hospitalizations,' 'Total apparent stimulant toxicity deaths,' 'Total stimulant-related poisoning hospitalizations,' 'Overall numbers', etc., were excluded to avoid data duplication and double counting.

6) An additional step of data transformation will be performed before running the predictive models.

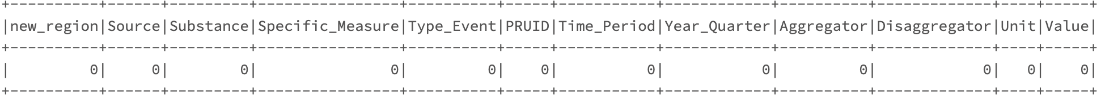
The data frame schema of the finalized version of the file looks as follow:

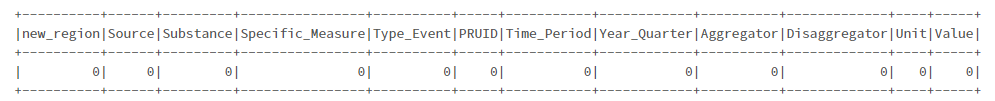
Text

Description automatically generated

**Figure 3.** It shows the dataframe schema for opioid dataset.

The final dataset consists of 936 rows with the number of opioid-related death and emergency services by province, gender, and age group and does not include any records with missing values.



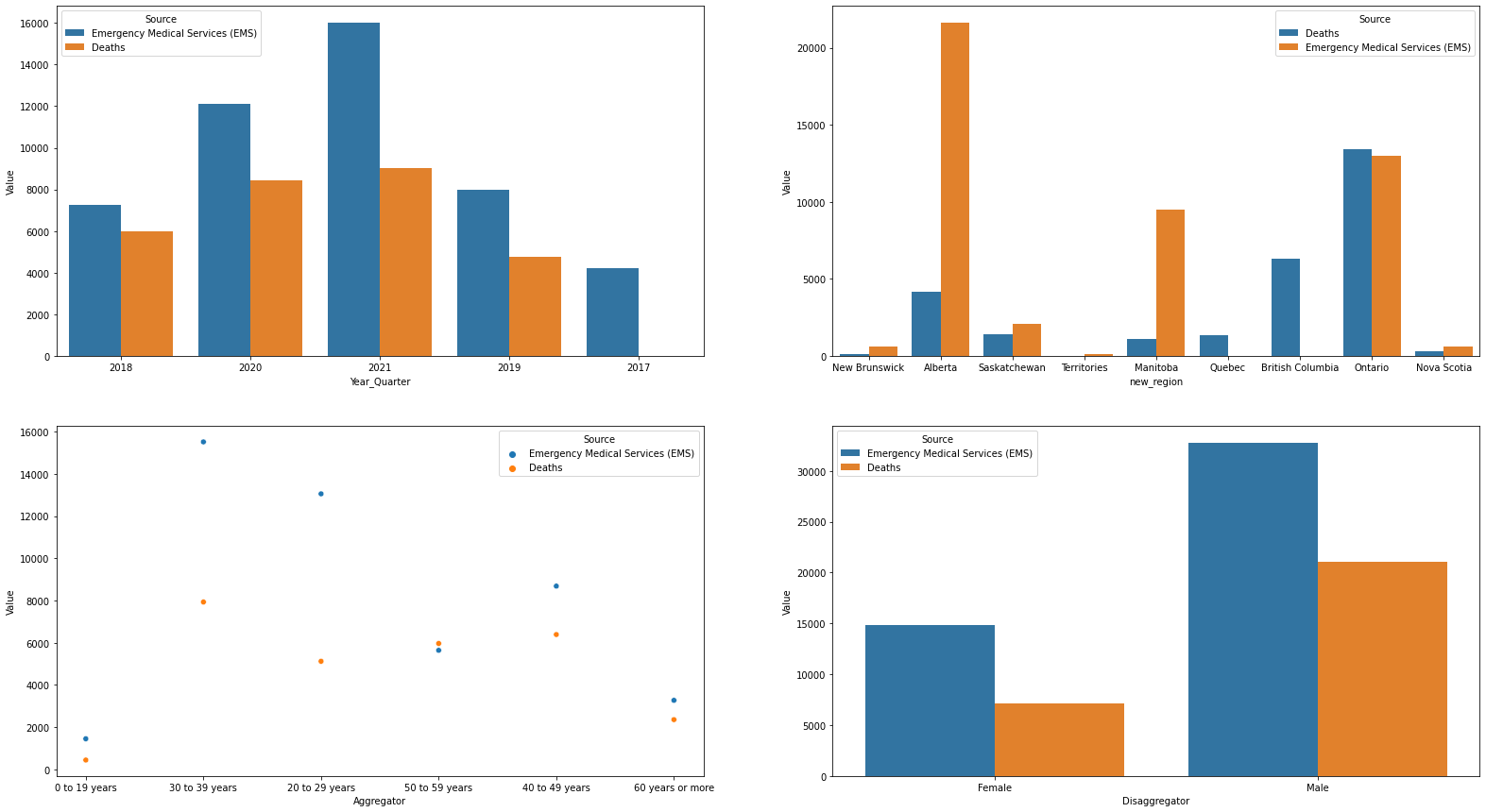


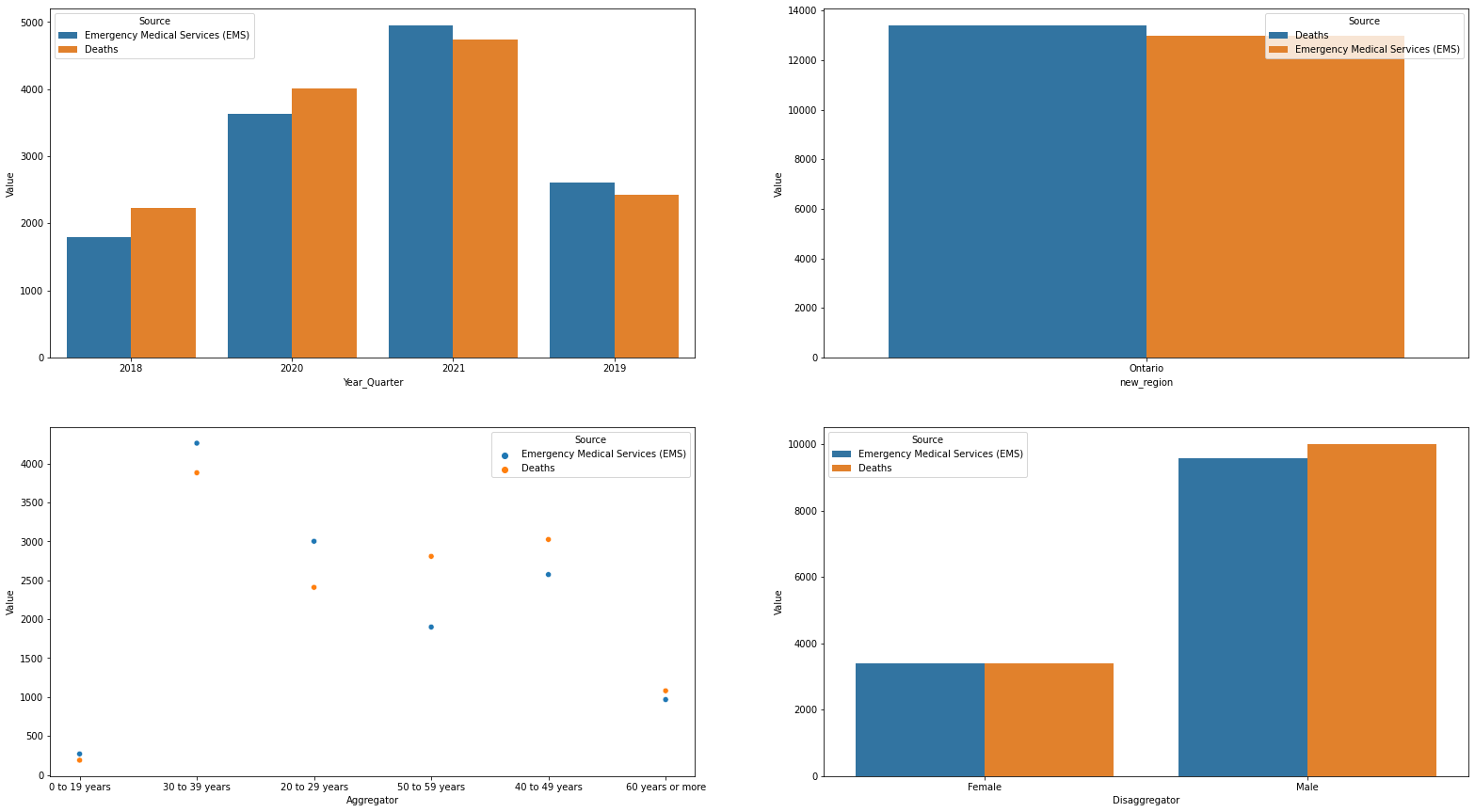
**Figure 4.** Data cleaning output.

* + 1. **Primary analysis/assessment and visualization/findings**

Overall, 71590 opioid/stimulant toxicity cases occurred in Canada in four years from 2018 to 2021 (Table A.1). The highest number of toxicity episodes occurred in Ontario and Alberta (Table A.2). 65% of Emergency Medical Services and 61% of death incidents happened during 2020 and 2021, which suggests a possible correlation between opioid/stimulant consumption and mental and physical outcomes of the COVID-19 pandemic. The highest absolute number of opioid toxicity episodes occurred among those who were 30-39 years old, followed by 20-29 years old. However, the highest proportion of opioid/stimulant toxicity deaths occurred among those 50 to 59 years old (Table A.3). Males comprise 71% of opioid toxicity cases with a 39% death rate (33% of death outcomes for females) (Table A.4).

In Canada, 39% of opioid/stimulant toxicity cases resulted in a death compared to 51% of cases with death outcomes in Ontario. Otherwise, Ontario follows the overall patterns (Figure 5, Figure 6).

**Figure 5. Outcome of opioid or stimulant toxicity by demographic characteristics (Canada)**



**Figure 6. Outcome of opioid or stimulant toxicity by demographic characteristics (Ontario)**

1. **Results**

**5.1. Data Preparation for Machine Learning**

The data preparation for the ML process included changing the data structure. The original dataset was structured as an aggregation of the total events (death of emergency services) by province/territory and demographic information. The PySpark explode transformation function was applied to the dataset, and the final version of the dataset included a new row for each element in an array so that this dataset can be used for modelling.

This analysis aimed to identify the predictors of opioid toxicity death for all Canadian provinces and territories and separately for Ontario; therefore, the next step of data preparation was creating a separate dataset for Ontario.

The third step of data preparation for machine learning included encoding string columns of labels to columns of numeric indices. The following columns were encoded: new\_region (Province/territory data), Source (Death or Emergency Medical Services), Substance (Opioid, Stimulants), Year\_Quarter (Year), Aggregator (Age group), and Disaggregation (Sex). The StringIndexer encoded labels in ascending order, with the most frequent labels indexed as 0.

**5.2 Model Development**

A pipeline was created to run the Transformer and Estimator stages in a particular sequence for machine learning.

The Source variable was used as a dependent variable (label variables), and the ML techniques were used to predict the outcome of opioid or stimulant related harm.

VectorAssembler was used to combine independent variables (new\_region\_indexer, Substance\_indexer, Year\_Quarter\_indexer, Aggregator\_indexer, Disaggregator\_indexer) into the vector format for further regression analysis.

The data was split into train (70%, n=50118) and test (30%, n=21472) datasets and three ML models were used to analyze data: Random Forest, Linear Regression, Logistic Regression.

**5.3 Machine Learning models**

In my research, I used three machine learning models to predict the outcome of opioid or stimulant toxicity: Linear Regression, Logistic Regression, and Random Forest.

Linear regression finds the best-fit line (i.e., linear relationships) between the dependent variable and independent variables. Linear regression is typically performed on a continuous dependent variable. The linear regression model consists of dependent variables (label) and explanatory variables (features vector) and is fitted on the vectorized training dataset. The model performance was evaluated using the testing dataset.

Logistic regression models the probability of an event (death outcome of opioid/stimulant toxicity in my research) based on the demographic characteristics of the individual. Logistic regression is a statistical technique that is primarily used to analyze binary dependent variables, and it aims to find the best weight coefficient for each independent variable.

The Random Forest algorithm constructs multiple samples of the training dataset, makes a prediction for each sample, and averages out those predictions to get the most accurate estimate of the output value. Random Forest algorithms can be used on both continuous and binary data.

For each model, I collected performance metrics, and the results are presented in Table 1.

**Table 1. ML model characteristics – All Canada**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Linear Regression Model** | **Logistic Regression Model** | **Random Forest** |
| Accuracy score | 0.746 | 0.725 | 0.810 |
| MSE | 0.253 | 0.274 | 0.189 |
| MAE | 0.253 | 0.274 | 0.189 |
| RMSE Squared | 0.503 | 0.523 | 0.435 |
| Explained Variance | 0.185 | 0.18 | 0.198 |

### Random Forest Algorithm showed the highest Accuracy Score and Explained Variance, and the lowest Mean Squared Error, Mean Absolute Error, and Root Mean Square Error. Place of residence, gender and age group are strong predictors of opioid/stimulant toxicity outcome.

### The next step is to check if I get comparable results when I test Ontario-only data using the same machine-learning models.

**Table 2. ML model characteristics – Ontario**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Linear Regression Model** | **Logistic Regression Model** | **Random Forest** |
| Accuracy score | 0.256 | 0.259 | 0.697 |
| MSE | 0.697 | 0.740 | 0.302 |
| MAE | 0.302 | 0.740 | 0.302 |
| RMSE Squared | 0.302 | 0.860 | 0.550 |
| Explained Variance | 0.209 | 0.324 | 0.256 |

### Like results for all of Canada, the Random Forest model showed the highest accuracy score and variance explained for Ontario only data.

**Conclusion**

Accuracy score is one of the most important characteristics of the ML learning as it shows the percentage of correct predictions made by a model; hence, I would suggest using the Random Forest model for further analysis of opioid/stimulant toxicity.

**Reference:**

1. [Lisa Belzak](https://pubmed.ncbi.nlm.nih.gov/?term=Belzak+L&cauthor_id=29911818), [Jessica Halverson](https://pubmed.ncbi.nlm.nih.gov/?term=Halverson+J&cauthor_id=29911818). The opioid crisis in Canada: a national perspective. Health Promot Chronic Dis Prev Can. 2018 Jun;38(6):224-233.
2. Higashikawa Y, Suzuki S. Studies on 1-(2-phenethyl)-4-(N-propionylanilino)piperidine (fentanyl) and its related compounds. VI. Structure-analgesic activity relationship for fentanyl, methyl-substituted fentanyls and other analogues. Forensic Toxicol. 2008;26(1):1-5. doi:10.1007/s11419-007-0039-1
3. <https://health-infobase.canada.ca/substance-related-harms/opioids-stimulants/technical-notes>

**Appendix**

**Table A.1**

|  |  |  |  |
| --- | --- | --- | --- |
| Year\_Quarter | Source | Canada | Ontario |
| 2018 | Emergency Medical Services (EMS) | 7269 | 1789 |
| 2018 | Deaths | 5983 | 2223 |
| 2019 | Emergency Medical Services (EMS) | 7990 | 2611 |
| 2019 | Deaths | 4768 | 2422 |
| 2020 | Emergency Medical Services (EMS) | 12108 | 3627 |
| 2020 | Deaths | 8439 | 4015 |
| 2021 | Emergency Medical Services (EMS) | 16010 | 4949 |
| 2021 | Deaths | 9023 | 4735 |

**Table A.2**

|  |  |  |
| --- | --- | --- |
| new\_region | Source | Value |
| Alberta | Emergency Medical Services (EMS) | 21632 |
| Alberta | Deaths | 4188 |
| British Columbia | Deaths | 6337 |
| Manitoba | Deaths | 1132 |
| Manitoba | Emergency Medical Services (EMS) | 9493 |
| New Brunswick | Deaths | 128 |
| New Brunswick | Emergency Medical Services (EMS) | 635 |
| Nova Scotia | Emergency Medical Services (EMS) | 630 |
| Nova Scotia | Deaths | 279 |
| Ontario | Deaths | 13395 |
| Ontario | Emergency Medical Services (EMS) | 12976 |
| Quebec | Deaths | 1363 |
| Saskatchewan | Deaths | 1388 |
| Saskatchewan | Emergency Medical Services (EMS) | 2094 |
| Territories | Emergency Medical Services (EMS) | 150 |
| Territories | Deaths | 3 |

**Table A.3**

|  |  |  |  |
| --- | --- | --- | --- |
| Aggregator | Source | Canada | Ontario |
| 0 to 19 years | Emergency Medical Services (EMS) | 1453 | 270 |
| 0 to 19 years | Deaths | 447 | 189 |
| 20 to 29 years | Emergency Medical Services (EMS) | 13051 | 3002 |
| 20 to 29 years | Deaths | 5122 | 2410 |
| 30 to 39 years | Emergency Medical Services (EMS) | 15510 | 4262 |
| 30 to 39 years | Deaths | 7929 | 3882 |
| 40 to 49 years | Emergency Medical Services (EMS) | 8682 | 2574 |
| 40 to 49 years | Deaths | 6389 | 3025 |
| 50 to 59 years | Emergency Medical Services (EMS) | 5641 | 1900 |
| 50 to 59 years | Deaths | 5966 | 2808 |
| 60 years or more | Deaths | 2360 | 1081 |
| 60 years or more | Emergency Medical Services (EMS) | 3273 | 968 |

**Table A.4**

|  |  |  |  |
| --- | --- | --- | --- |
| Disaggregator | Source | Canada | Ontario |
| Male | Deaths | 21048 | 10004 |
| Female | Deaths | 7165 | 3391 |
| Female | Emergency Medical Services (EMS) | 14823 | 3396 |
| Male | Emergency Medical Services (EMS) | 32787 | 9580 |