An Analysis of California Wildfires from 2013 to 2020 with PySpark

BIG DATA ANALYTICS AND DATA VISUALISATION 7153CEM

Faculty of Engineering, Environment and Computing
MSc Data Science
Coventry University
Module Leader: Dr. Marwan Fuad

Alexandra Espialidou
Student ID: 13047171

Contents

Word Count: 3092

Abstract

This report presents a comprehensive analysis of the California wildfires dataset providing valuable insights into the occurrences, causes and impacts of wildfires in the state. Through thorough data analysis, including exploratory data visualization and regression analysis, this study aims to uncover trends, draw meaningful conclusions, and provide insights.

Introduction

This paper's main purpose is to analyze the California wildfire dataset and derive valuable insights. Wildfires are deadly and pose a recurring and pressing environmental concern. They are catastrophic to various communities, ecosystems and infrastructure. A report by the California Natural Resources Agency in 2018[1] revealed that should current global warming trends persist, wildfires in California will become even more frequent and prolonged. California, the second-largest grassland-covered state after Alaska, encompasses roughly one-third of its area in forests and grasslands. Moreover, the west coast of the U.S.A., where California is located, is characterized by a hot and dry climate. Prolonged droughts and high temperatures are taking advantage of the abundant forests rendering the region highly susceptible to wildfires that can persist for days. This underscores the significance of analyzing this dataset, the insights from which can prove highly valuable. The dataset contains various attributes, including fire incident details, location information, fire characteristics, response resources, impact and damage. By examining this information, the most endangered counties can be identified, along with the years marked by the most severe fires. Furthermore, insights into response times can be gleaned by examining available resources and burned acreage. The aim of this paper is to find some reasoning as to how can we decrease the occurrence of such incidents, minimize their impact, and improve the response. This analysis was conducted with the help of PySpark, a data processing library within the Apache Spark ecosystem and Jupyter Notebook. The visualizations were carried out by Tableau.

The dataset

The dataset used for this paper was obtained from Kaggle.com. It was created by a person with the alias 'ARES' in 2019. The dataset in named *California_Fire_Incidents.csv* and contains he data of over 1600 wildfires that have occurred in California between 2013 and 2020. With 1636 instances/rows and 40 attributes it paints a very detailed picture of how devastating these fires were. Each instance in the dataset corresponds to the comprehensive information of a fire incident in California. *F8 point* (which can be found in the appendix along with the totality of the code and its output) presents all the attributes and their type.

As mentioned, the dataset consists of 40 columns. These include: links, name and county IDs, location descriptions, incident statements, acres burned, personnel and equipment used, location details, names of the fires, structures damaged or threatened and start and extinguished dates. In *F3 point* a screenshot from the data in the csv file can be seen.

Data processing

When the dataset was imported with `df = spark.read.csv(csv_path, header=True, inferSchema=True)` the datatype of all the columns was automatically converted to string. The next step was to convert it back to its original type. As you can see in *F8 point* the dataset consists of 16 floats, 1 date, 3 timestamps, 6 boolean and 14 string data. The most important attributes to this analysis are presented in *Table 1* below:

Table 1

Attribute	Type	Description
AcresBurned	float	Number of Acres Burned by fire incidents
Counties	string	County where each fire started
CrewsInvolved	float	Number of fire crews involved in the incident
Dozers	float	Number of bulldozers involved in the incident
Engines	float	Number of engines involved in the incident

[1] https://www.theguardian.com/world/ng-interactive/2018/sep/20/why-are-california-wildfires-so-bad-interactive

Extinguished	timestamp	Extinguished date of the fire
Fatalities	float	Number of people that lost their lives
Helicopters	float	Number of helicopters assigned
Injuries	float	Number of injured personnel
Latitude	float	Latitude of the incident
Longitude	float	Longitude of the incident
MajorIncident	boolean	Whether it was considered a major incident or not
Name	string	Name of the wildfire
PersonnelInvolved	float	Number of CalFire personnel involved
Started	timestamp	Start date of the fire incident
WaterTenders	float	Number of trucks carrying water used for the fire incident

The type of the dataset was checked and it is an instance of the PySpark DataFrame class (*F7 point*). This means that it is structured and enables distributed computation across a cluster of machines. For this project there is one master node (`local[*]`) as can be seen in *F4 point*.

The next step in the data processing phase involves the examination of Null values within the dataset. The dataset does not provide explicit clarification regarding whether these Null values correspond to missing data or instances where the recorded numerical value is indeed zero, so for the sake of the analysis, all the Null values will be converted to the numerical value 0, as can be seen in *F9 point* with the help of *pyspark.sql.functions*. Moreover, the dataset was checked for duplicates and it does not contain any. Unrelated data from 1969 were deleted.as can be seen in *F10 point*. The final dataset that is ready to use contains 1458 instances and 40 attributes.

The Describe() function gives an insight in some key metrics of the dataset. As can be seen in *F11 point,* the summary of count, mean, standard deviation, minimum and maximum for the 3 columns mentioned above are presented.

To begin with, the average acres that were burned was approximately 1,981 acres, with a high standard deviation of around 11090 acres. The minimum indicates that there are cases with 0 acres burned, which are likely fires that were controlled or just small incidents. On the other hand, the maximum was 257,314 acres indicating large-scale wildfires.
As for latitude and longitude, the values indicate a general centering within California, as the dataset suggests. However, the high standard deviations in both latitude and longitude emphasize the diverse geographic distribution of wildfires across the state.

Another interesting observation is regarding the summary metrics of PersonnelInvolved, which is CalFire[2] personnel and crews involved. The average number of personnel involved in firefighting operations is approximately 365, with a standard deviation of about 795.64. The minimum value of 0 indicates that there might be instances where no personnel were involved, potentially indicating a controlled or minor incident or a lack of data for some instances. The maximum value of 5636 emphasize the substantial capacity mobilized for managing large-scale wildfire incidents. As for crews, their range varies from a minimum of 1 crew to a maximum of 82 crews.
To conclude, these metrics give valuable insights into the scale and resources allocated to firefighting efforts during California wildfires. These numbers provide a sense of the magnitude of human resources dedicated to combat these incidents. It is important to note that a crew consists of multiple people, so the number 10 does not mean few people.

Methodology

In order to facilitate the analysis of the California wildfires dataset, Pyspark, the Python API for Apache, was used. The installation process was simple. As documented in *F0 point*, the installation of Apache Spark was initiated through the Command Prompt, employing the command: 'pip install PySpark'.

[2] CalFire is the California Department of Forestry and Fire Protection, a state agency that is responsible for fire protection, wildfire response, and management of natural resources. (fire.ca.gov)

Within Jupyter Notebook (*F1 point*) the PySpark library and SparkSession were imported. Finally, a Spark session was configured with 'Coursework' as the name of the app, as seen in *F4 point.*

All figures built for this analysis were created using the professional edition 2023.2.0 (20232.23.0611.2007) of Tableau Desktop. (*F2 point*)

Lastly, the results of this analysis were acquired with the use of HP 15s - eq2035nv (laptop), equipped with AMD Ryzen 5 4500U CPU with Radeon Graphics with 2.38 GHz and 8.00 GB RAM.

Regarding the systematic approach of the analysis, the dataset was analyzed using the PySpark library. Prior to the analysis the dataset underwent thorough preprocessing to ensure its quality and consistency, by handling Null values and duplicates. In the data analysis process, exploratory analysis encompassed summary statistics and visualizations to unveil patterns and relationships within the data. Visualizations, created in Tableau including geographical heat maps, bar graphs and line graphs were generated to present a clear portrayal of the dataset's characteristics. Lastly, 2 regression models and 1 classification model were built with the help of VectorAssembler, RandomForestRegressor, LinearRegression and LogisticRegression from the pyspark.ml library.

Exploratory Data Analysis

General Observations

In *F12 point* several notable observations about the dataset can be seen. Over the period from 2013 to 2020, 4,961,377 aces were burned by fire incidents, a total of 211 reported injuries occurred and 40,185 personnel were involved in their extinguishment. Additionally, it is noteworthy that the year 2017 witnessed the highest count of damaged structures, reaching a total of 3,408. Similarly, in the following year, 2018, the dataset recorded the highest numbers of both damaged structures (26,855) and threatened structures (7,285).

Top 10 Affected Counties

In *F13 point,* the top 10 affected counties by Acres Burned and Injuries can be seen. The county with the biggest fires (most acres burned) is **Lake County** with 582,784 acres and the county with the most injuries is **Shasta** with 55.
In *F14 point*, the top 10 affected counties by number of fires can be seen. The county with the most fire incidents is **Riverside** with 140.

Deadliest Year

In *F15 point* the graphs show the density of started fires and acres burned in each year. It is evident that the year with the most fires is **2017** with 437 fires. However, the deadliest year (with the most aces burned) is **2018** with 3,358,004 acres.
In *F16 point*, the burned acres and injuries per year can be seen together. The deadliest year, regarding acres burned, is indeed **2018**, with the numbers starting to advance from 2016. However, it is interesting to see that the injuries were at their highest in **2014** with 137 and just 28 in 2018. As for fatalities, **2018** had the majorities of fatalities with 102, as demonstrated in *F17 point*.
As for the deadliest month, from *F18 point* it is clear that the **summer months (June, July, August)** are the most susceptible for a fire incident. In June there are 319 incidents, in July 415 incidents and in August 282 incidents. The month with the least fire incidents is March with just 6 fire incidents.

Spatial Visualization

In *F19 point*, the 2 geographical heat maps show the location of fire incidents from 2013 to 2020. The first map gives a general representation of the totality of the fire incidents. The second map shows the 20 largest (by acres burned) fires. The most important observation to be made extracted from these graphs is that the fires are happening in deep forested areas. The biggest one took place between Santa

Barbara and Los Angeles, where 563,786 acres were burned in 2017. These maps help visualize the incidents and give comprehensive and multidimensional perspective to the incidents.

Personnel and Equipment

F20 point shows a graph with the personnel that were involved in the extinguishment of the fires and the Water Tankers that were used. It is demonstrated that these numbers were proportional. Another interesting observation is that in 2018, which was the deadliest year for fire incidents, the second highest numbers can be seen in both attributes with 348 in Water Tenders and 13,768 personnel. This indicates the substantial demand for personnel and equipment.

In *F21 point*, the graph shows the usage of equipment (bulldozers, Water Tenders, Engines). The majority of them was used in 2013 and 2018. 2018 was indeed a very bad year, incident wise, but 2013, according to the dataset, does not show very high indicators. Perhaps the fact that the personnel and equipment were at a high, helped limit the damages.

Regression models and Evaluation

Ridge Linear Regression

Before starting this process, the dataset was preprocessed, as can be seen in the dataset section. However, the column 'Started' needed to be converted to 'float' (from 'timestamp') in this part, for the model to work.

In this step a Ridge regression model was initially implemented to address multicollinearity issues present in the dataset. The dataset has data that might be correlated like location attributes (latitude and longitude) and resources (AirTankers, WaterTenders, Helicopters, PersonnelInvolved) which can negatively impact the model, leading to predicted values that are significantly distant from the actual values[3]. For the implementation, the PySpark library's LinearRegression module was imported, along with VectorAssembler and RegressionEvaluator. The selected features were location attributes (latitude, longitude), date where the fire started, number of air tankers, number of water tankers, helicopters and available personnel. These features were incorporated into the VectorAssembler, transformed to create the feature vectors required by the model and divided into training and testing sets. Subsequently, they were trained with the LinearRegression function. In order to generate predictions, the testing set was utilized. Detailed code and prediction results can be found in *F22 point.*

Evaluation of the Ridge regression model

The Root Mean Squared Error (RMSE) for this model is 8967.552848686488. This value implies that the model's predictions deviate from the actual values by around 8767.53 acres. The significance of this deviation depends on the context and unit of measurement. In this case, where the unit is acres, 8767.53 acres translate to roughly 35.4 kilometers. To conclude, this model can predict the acres that will be burned, based on known latitude, longitude and resources with an error of approximately 35.4 kilometers.

Random Forests model

Given that an error of 35.4 kilometers signifies a considerable level of inaccuracy, an alternative approach was pursued by constructing a secondary regression model - specifically, a Random Forest model. A Random Forest model uses multiple models, which are trained over the same data[4]. The average of the results of all the models is calculated, giving a more accurate prediction. The selected features remained the same as the previous model. The model was split, fitted and evaluated with the RandomForestRegressor and the RegressionEvaluator as detailed in *F3 point*.

Evaluation of the Random Forests model

The value of RMSE of the new model was 8757.530616400194, indicating again that the model is off by approximately the same number of values as before. These consistent results suggest that the selected attributes, despite satisfying the criteria, might not be ideally suited for this particular approach.

[3] https://www.mygreatlearning.com/blog/what-is-ridge-regression/

[4] https://towardsdatascience.com/random-forest-regression-5f605132d19d

Logistic Regression model

With the objective in mind, a classification model was developed in order to try and identify different aspects of the dataset. This time a logistic regression model was constructed aiming to classify fire incidents as “Major” occurrences. Logistic regression is a supervised statistical method ideal for binary classification tasks. It maps the selected features and corresponding to the possibility of being 1 or 0 it groups data together.

In *F24 point* the count of TRUE and FALSE values of major incidents are presented, where True indicates a major incident. There are 383 True values and 1253 False values. The model's focus is to predict, based on this distribution, whether a fire incident qualifies as "Major. The selected features were burned acres, injuries and available personnel and the target column was MajorIncident. To facilitate classification, the target column was transformed: TRUE values were assigned the label 1, while FALSE values were designated as 0, as can be seen in *F25a point*. The features were incorporated into the VectorAssembler, split and fitted. The logistic regression model was established employing the LogisticRegression function. A workflow was created through the formulation of a pipeline and the model was trained. The model was evaluated with the MulticlassClassificationEvaluator library.

Evaluation of the Logistic Regression model

The classification report(*F25b point*) presents the results of the logistic regression model. Precision is the measure of the proportion of True positives out of all the predicted instances. Recall, also referred to as sensitivity is the number of all True positive instances out of all the actual positive instances. F1-Score is the mean of precision and recall. In this model the numbers are high, suggesting a good model. Precision stands at 0.8916, recall is 0.8797 and F1-Score is 0.8594. These figures collectively suggest the model's strength and its capability in accurately classifying fire incidents as "Major" or not.

Result Discussion and Conclusion

The experimental data analysis results shed light on crucial insights regarding the behavior and attributes of California wildfires. Over the 2013-2020 period, 4,961,377 acres were burned, and 211 injuries occurred, underlining the substantial environmental and societal impact. Notably, 2017 and 2018 saw the highest counts of damaged and threatened structures, emphasizing the need for effective mitigation strategies. Lake County experienced the most extensive damage, with 582,784 acres burned, while Riverside County recorded the highest fire frequency with 140 incidents. Although 2017 witnessed the highest number of fires, 2018 marked the most devastating year, encompassing a staggering 3,358,004 acres burned. Geographical analysis revealed that wildfires mainly occurred in forested areas, with the largest one occurring between Santa Barbara and Los Angeles in 2017. Personnel and equipment deployment indicated a resource-intensive approach during critical years.

Regression and classification methods like Ridge regression, Random Forests, and Logistic Regression explored correlations between attributes and major incidents. The Ridge regression model aimed to address multicollinearity concerns showing a predictive error of around 35.4 kilometers, provided a baseline for further improvement. Random Forests model delivered consistent results, indicating that a more refined selection of features might be needed for accurate predictions. The introduction of the Logistic Regression model added a classification perspective, successfully distinguishing "Major" incidents based on the distribution of selected attributes.

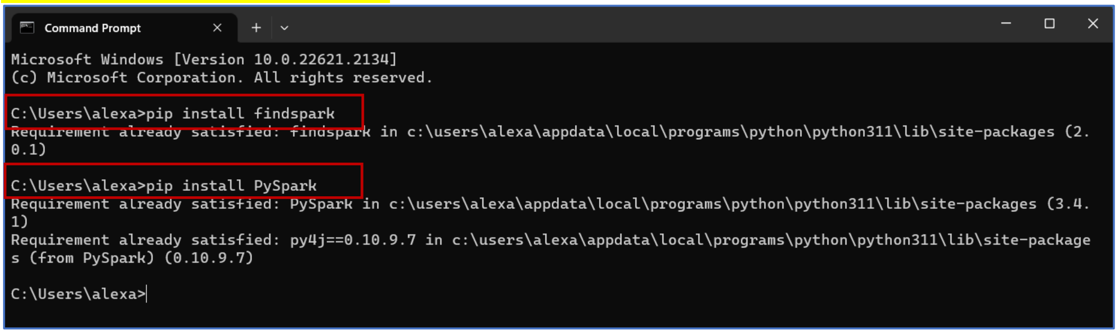
Social Impact

This study has a significant impact on addressing the social impact of wildfires. By analyzing and predicting the factors contributing to wildfire occurrences, the study provides valuable information for policymakers, emergency responders, and local communities in California and other states or countries with similar conditions for firefighting efforts and proactive disaster management. Moreover, the classification model's ability to predict major incidents can enhance early response strategies and help mitigate the potential destruction caused by large-scale fires. The findings not only contribute to the scientific understanding of wildfire dynamics but also have direct implications for public safety and community resilience. The paper's insights can assist in forming more informed policies, enhancing public awareness campaigns and show the impact these fire incidents have.

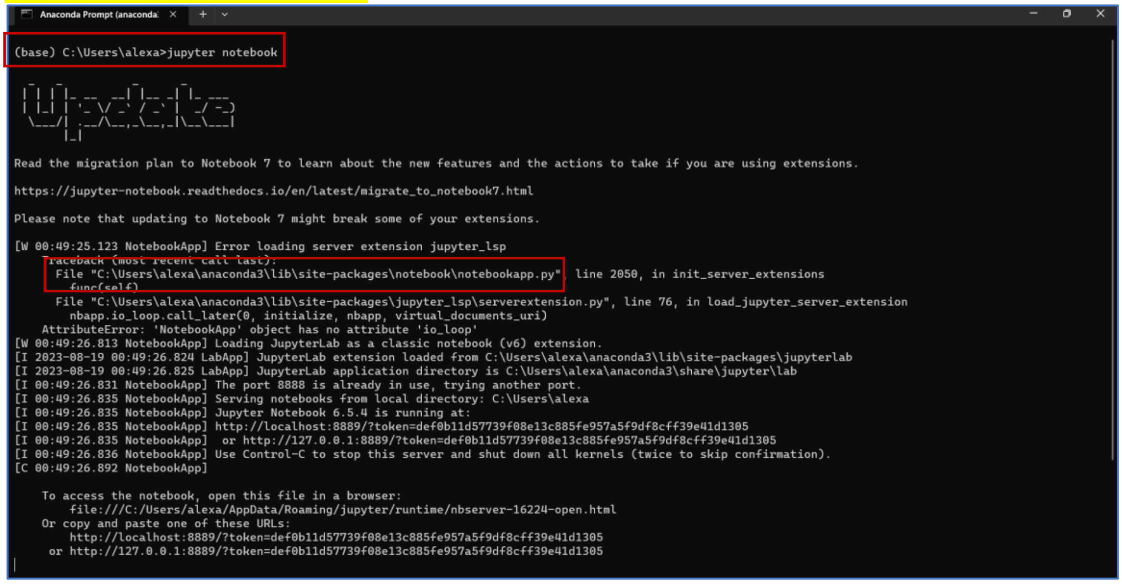
Ultimately, this analysis showcases how a data-driven analysis can have real-world applications leading to a better disaster anticipation as global warming keeps escalating.

Appendix

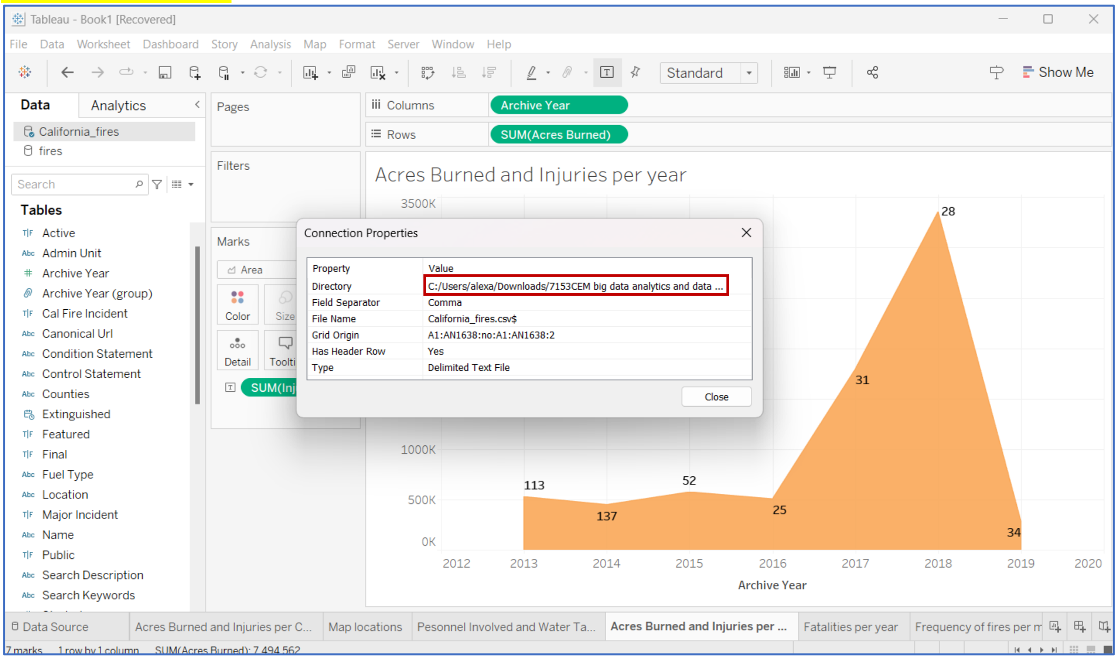
F0 point: PySpark installation



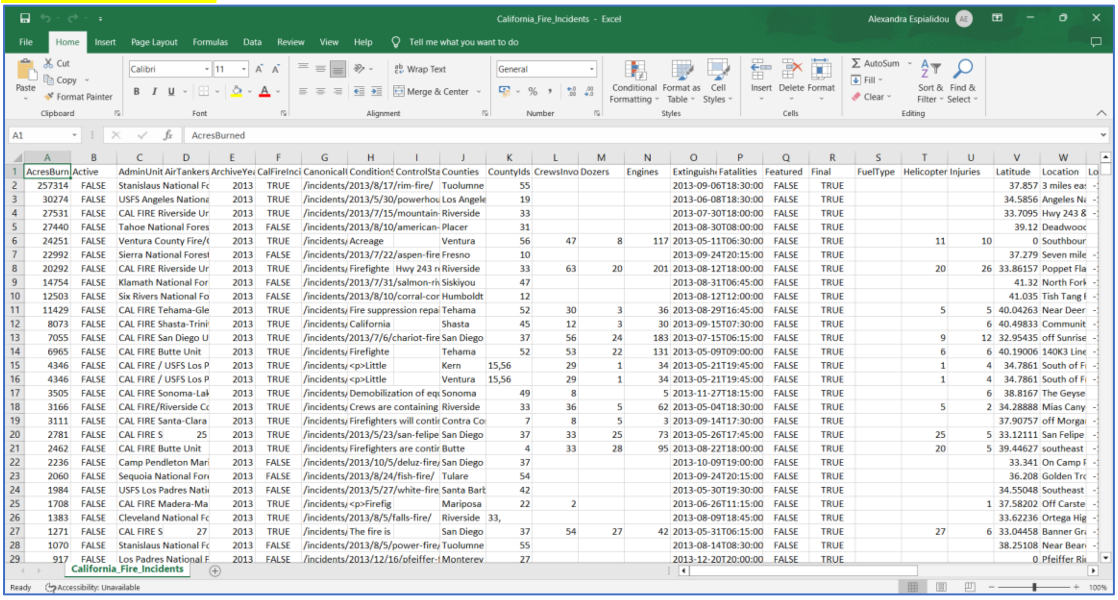
F1 point: Jupyter Notebook



F2 point: Tableau



F3 point: csv file



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
1	AcresBurn	Active	AdminUnit	AirTankers	ArchiveYea	CalFireInci	CanonicalU	Condition:	ControlSta	Counties	CountyIds	CrewsInvo	Dozers	Engines	Extinguish	Fatalities	Featured	Final	FuelType	Helicopter	Injuries	Latitude	Location	Lo
2	257314	FALSE	Stanislaus National Fc		2013	TRUE	/incidents/2013/8/17/rim-fire/			Tuolumne	55				2013-09-06T18:30:00		FALSE	TRUE				37.857	3 miles ea:	-
3	30274	FALSE	USFS Angeles Nationa		2013	TRUE	/incidents/2013/5/30/powerhou			Los Angele	19				2013-06-08T18:30:00		FALSE	TRUE				34.5856	Angeles Na	-
4	27531	FALSE	CAL FIRE Riverside Ur		2013	TRUE	/incidents/2013/7/15/mountain-			Riverside	33				2013-07-30T18:00:00		FALSE	TRUE				33.7095	Hwy 243 &	-
5	27440	FALSE	Tahoe National Fores		2013	FALSE	/incidents/2013/8/10/american-			Placer	31				2013-08-30T08:00:00		FALSE	TRUE				39.12	Deadwooc	
6	24251	FALSE	Ventura County Fire/(		2013	TRUE	/incidents/	Acreage		Ventura	56	47	8	117	2013-05-11T06:30:00		FALSE	TRUE		11	10	0	Southbour	
7	22992	FALSE	Sierra National Forest		2013	FALSE	/incidents/2013/7/22/aspen-fire			Fresno	10				2013-09-24T20:15:00		FALSE	TRUE				37.279	Seven mile	-
8	20292	FALSE	CAL FIRE Riverside Ur		2013	TRUE	/incidents/	Firefighte	Hwy 243 r	Riverside	33	63	20	201	2013-08-12T18:00:00		FALSE	TRUE		20	26	33.86157	Poppet Fla	-
9	14754	FALSE	Klamath National For		2013	FALSE	/incidents/2013/7/31/salmon-ri			Siskiyou	47				2013-08-31T06:45:00		FALSE	TRUE				41.32	North Fork	-
10	12503	FALSE	Six Rivers National Fo		2013	FALSE	/incidents/2013/8/10/corral-cor			Humboldt	12				2013-08-12T12:00:00		FALSE	TRUE				41.035	Tish Tang I	-
11	11429	FALSE	CAL FIRE Tehama-Gle		2013	TRUE	/incidents/	Fire suppression repai		Tehama	52	30	3	36	2013-08-29T16:45:00		FALSE	TRUE		5	5	40.04263	Near Deer	-
12	8073	FALSE	CAL FIRE Shasta-Trinit		2013	TRUE	/incidents/	California		Shasta	45	12	3	30	2013-09-15T07:30:00		FALSE	TRUE			6	40.49833	Communit	-
13	7055	FALSE	CAL FIRE San Diego U		2013	TRUE	/incidents/2013/7/6/chariot-fire			San Diego	37	56	24	183	2013-07-15T06:15:00		FALSE	TRUE		9	12	32.95435	off Sunrise	-
14	6965	FALSE	CAL FIRE Butte Unit		2013	TRUE	/incidents/	Firefighte		Tehama	52	53	22	131	2013-05-09T09:00:00		FALSE	TRUE		6	6	40.19006	140K3 Line	-
15	4346	FALSE	CAL FIRE / USFS Los P		2013	TRUE	/incidents/	<p>Little		Kern	15,56	29	1	34	2013-05-21T19:45:00		FALSE	TRUE		1	4	34.7861	South of Fi	-
16	4346	FALSE	CAL FIRE / USFS Los P		2013	TRUE	/incidents/	<p>Little		Ventura	15,56	29	1	34	2013-05-21T19:45:00		FALSE	TRUE		1	4	34.7861	South of Fi	-
17	3505	FALSE	CAL FIRE Sonoma-Lak		2013	TRUE	/incidents/	Demobilization of equ		Sonoma	49	8		5	2013-11-27T18:15:00		FALSE	TRUE			6	38.8167	The Geyse	
18	3166	FALSE	CAL FIRE/Riverside Cc		2013	TRUE	/incidents/	Crews are containing		Riverside	33	36	5	62	2013-05-04T18:30:00		FALSE	TRUE		5	2	34.28888	Mias Cany	-
19	3111	FALSE	CAL FIRE Santa-Clara		2013	TRUE	/incidents/	Firefighters will contir		Contra Co:	7	8	5	3	2013-09-14T17:30:00		FALSE	TRUE				37.90757	off Morgai	-
20	2781	FALSE	CAL FIRE S	25	2013	TRUE	/incidents/2013/5/23/san-felipe			San Diego	37	33	25	73	2013-05-26T17:45:00		FALSE	TRUE		25	5	33.12111	San Felipe	-
21	2462	FALSE	CAL FIRE Butte Unit		2013	TRUE	/incidents/	Firefighters are contir		Butte	4	33	28	95	2013-08-22T18:00:00		FALSE	TRUE		20	5	39.44627	southeast	-
22	2236	FALSE	Camp Pendleton Mari		2013	FALSE	/incidents/2013/10/5/deluz-fire/			San Diego	37				2013-10-09T19:00:00		FALSE	TRUE				33.341	On Camp I	-
23	2060	FALSE	Sequoia National For		2013	FALSE	/incidents/2013/8/24/fish-fire/			Tulare	54				2013-09-24T20:15:00		FALSE	TRUE				36.208	Golden Trc	-
24	1984	FALSE	USFS Los Padres Nati		2013	FALSE	/incidents/2013/5/27/white-fire			Santa Bart	42				2013-05-30T19:30:00		FALSE	TRUE				34.55048	Southeast	-
25	1708	FALSE	CAL FIRE Madera-Ma		2013	TRUE	/incidents/	<p>Firefig		Mariposa	22	2			2013-06-26T11:15:00		FALSE	TRUE			1	37.58202	Off Carste	-
26	1383	FALSE	Cleveland National Fc		2013	TRUE	/incidents/2013/8/5/falls-fire/			Riverside	33,				2013-08-09T18:45:00		FALSE	TRUE				33.62236	Ortega Hig	-
27	1271	FALSE	CAL FIRE S	27	2013	TRUE	/incidents/	The fire is		San Diego	37	54	27	42	2013-05-31T06:15:00		FALSE	TRUE		27	6	33.04458	Banner Gra	-
28	1070	FALSE	Stanislaus National Fc		2013	FALSE	/incidents/2013/8/5/power-fire/			Tuolumne	55				2013-08-14T08:30:00		FALSE	TRUE				38.25108	Near Bear	-
29	917	FALSE	Los Padres National F		2013	FALSE	/incidents/2013/12/16/pfeiffer-f			Monterey	27				2013-12-20T20:00:00		FALSE	TRUE				0	Pfeiffer Ri	-

California_Fire_Incidents

F4 point: Imports

```
!pip install pyspark
import pyspark
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, to_date
from pyspark.sql.functions import when, count, col
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.regression import RandomForestRegressor
from pyspark.ml.evaluation import RegressionEvaluator
spark = SparkSession.builder.appName('Coursework').getOrCreate()
spark
```

SparkSession - in-memory
SparkContext

Spark UI

Version

`v3.4.1`

Master

`local[*]`

AppName

`Coursework`

F5 point: Import dataset and set max rows and columns for better display

```
csv_path = "California_Fire_Incident.csv"
df = spark.read.csv(csv_path, header=True, inferSchema=True)
spark.conf.set("spark.sql.repl.eagerEval.enabled", True)
spark.conf.set("spark.sql.repl.eagerEval.maxNumRows", 1637) #all the rows of the dataset are 1636
spark.conf.set("spark.sql.repl.eagerEval.truncate", 100)
```

F6 point: Print the dataset

```
#The dataset is too long so it will be shown separately in 7 parts
df.select(['AcresBurned','Active',
'AdminUnit','AirTankers','ArchiveYear','CalFireIncident','CanonicalUrl']).sh
ow()
df.select(['ConditionStatement','ControlStatement','Counties','CountyIds','C
rewsInvolved','Dozers','Engines']).show()
df.select(['Extinguished','Fatalities','Featured','Final','FuelType','Helico
pters','Injuries']).show()
df.select(['Latitude','Location','Longitude','MajorIncident','Name','Percent
Contained']).show()
df.select(['PersonnelInvolved','Public','SearchDescription','SearchKeywords'
,'Started','Status']).show()
df.select(['StructuresDamaged','StructuresDestroyed','StructuresEvacuated','
StructuresThreatened','UniqueId','Updated']).show()
df.select(['WaterTenders']).show()
```

```
+--------------------+--------------------+--------------------+----------+------------+---------------+--------------------+
|         AcresBurned|              Active|           AdminUnit|AirTankers| ArchiveYear|CalFireIncident|        CanonicalUrl|
+--------------------+--------------------+--------------------+----------+------------+---------------+--------------------+
|              257314|               False|Stanislaus Nation...|      null|        2013|           True|/incidents/2013/8...|
|               30274|               False|USFS Angeles Nati...|      null|        2013|           True|/incidents/2013/5...|
|                267"|2013-05-30T15:28:00Z|           Finalized|      null|        null|           null|                null|
|               27531|               False|CAL FIRE Riversid...|      null|        2013|           True|/incidents/2013/7...|
|               27440|               False|Tahoe National Fo...|      null|        2013|          False|/incidents/2013/8...|
|               24251|               False|Ventura County Fi...|      null|        2013|           True|/incidents/2013/5...|
|<p>Continue to mo...| fire damage insp...| and suppression ...|      null|        null|           null|                null|
|",,Ventura,56,47,...|          Camarillo"|                 0.0|      True|Springs Fire|            100|                2167|
|               22992|               False|Sierra National F...|      null|        2013|          False|/incidents/2013/7...|
|               20292|               False|CAL FIRE Riversid...|      null|        2013|           True|/incidents/2013/8...|
|Command of the in...|                null|                null|      null|        null|           null|                null|
|closed. For quest...| please the Silen...|Hwy 243 remains c...| Riverside|          33|             63|                  20|
|               14754|               False|Klamath National ...|      null|        2013|          False|/incidents/2013/7...|
|               12503|               False|Six Rivers Nation...|      null|        2013|          False|/incidents/2013/8...|
|               11429|               False|CAL FIRE Tehama-G...|      null|        2013|           True|/incidents/2013/8...|
|                8073|               False|CAL FIRE Shasta-T...|      null|        2013|           True|/incidents/2013/9...|
|<p>All evacuation...|                null|              Shasta|        45|          12|              3|                  30|
|                7055|               False|CAL FIRE San Dieg...|      null|        2013|           True|/incidents/2013/7...|
|                6965|               False| CAL FIRE Butte Unit|      null|        2013|           True|/incidents/2013/5...|
|lines for the nex...| timber slash pil...| remain hot and a...|      null|        null|           null|                null|
+--------------------+--------------------+--------------------+----------+------------+---------------+--------------------+
only showing top 20 rows

+--------------------+--------------------+--------------------+--------------------+-------------+------+-------+
|  ConditionStatement|    ControlStatement|            Counties|           CountyIds|CrewsInvolved|Dozers|Engines|
+--------------------+--------------------+--------------------+--------------------+-------------+------+-------+
|                null|                null|            Tuolumne|                  55|         null|  null|   null|
|                null|                null|         Los Angeles|                  19|         null|  null|   null|
|bf37805e-1cc2-420...|2013-06-08T18:30:00Z|                null|                null|         null|  null|   null|
|                null|                null|           Riverside|                  33|         null|  null|   null|
|                null|                null|              Placer|                  31|         null|  null|   null|
```

```
|Acreage has been ...|                null|                null|                null|        null|  null|   null|
|                null|                null|                null|                null|        null|  null|   null|
|                True|The Springs Fire ...|Springs Fire, May...|2013-05-02T07:01:00Z|   Finalized|     6|     10|
|                null|                null|              Fresno|                  10|        null|  null|   null|
|Firefighters clos...|                null|                null|                null|        null|  null|   null|
|                null|                null|                null|                null|        null|  null|   null|
|                 201|2013-08-12T18:00:00Z|                null|               False|        True|  null|     20|
|                null|                null|            Siskiyou|                  47|        null|  null|   null|
|                null|                null|            Humboldt|                  12|        null|  null|   null|
|Fire suppression ...|                null|              Tehama|                  52|          30|     3|     36|
|California Incide...|                null|                null|                null|        null|  null|   null|
|2013-09-15T07:30:00Z|                null|               False|                True|        null|  null|      6|
|                null|                null|           San Diego|                  37|          56|    24|    183|
|Firefighters cont...|                null|                null|                null|        null|  null|   null|
|                null|                null|                null|                null|        null|  null|   null|
+--------------------+--------------------+--------------------+--------------------+------------+------+-------+
only showing top 20 rows

+--------------------+--------------------+--------------------+--------------------+-----------+-----------+--------+
|        Extinguished|          Fatalities|            Featured|               Final|   FuelType|Helicopters|Injuries|
+--------------------+--------------------+--------------------+--------------------+-----------+-----------+--------+
|2013-09-06T18:30:00Z|                null|               False|                True|       null|       null|    null|
|2013-06-08T18:30:00Z|                null|               False|                True|       null|       null|    null|
|                null|                null|                null|                null|       null|       null|    null|
|2013-07-30T18:00:00Z|                null|               False|                True|       null|       null|    null|
|2013-08-30T08:00:00Z|                null|               False|                True|       null|       null|    null|
|                null|                null|                null|                null|       null|       null|    null|
|                null|                null|                null|                null|       null|       null|    null|
|                null|                null|46731fb8-3350-492...|2013-05-11T06:30:00Z|         11|       null|    null|
|2013-09-24T20:15:00Z|                null|               False|                True|       null|       null|    null|
|                null|                null|                null|                null|       null|       null|    null|
|                null|                null|                null|                null|       null|       null|    null|
|                  26|            33.86157|Poppet Flats Rd n...|          -116.90427|       True|Silver Fire|     100|
|2013-08-31T06:45:00Z|                null|               False|                True|       null|       null|    null|
|2013-08-12T12:00:00Z|                null|               False|                True|       null|       null|    null|
|2013-08-29T16:45:00Z|                null|               False|                True|       null|          5|       5|
|                null|                null|                null|                null|       null|       null|    null|
|           40.498332|Community of Igo,...|         -122.535496|                True|Clover Fire|        100|     342|
|2013-07-15T06:15:00Z|                null|               False|                True|       null|          9|      12|
|                null|                null|                null|                null|       null|       null|    null|
|                null|                null|                null|                null|       null|       null|    null|
+--------------------+--------------------+--------------------+--------------------+-----------+-----------+--------+
only showing top 20 rows

+---------+--------------------+--------------------+--------------------+--------------------+----------------+
| Latitude|            Location|           Longitude|       MajorIncident|                Name|PercentContained|
+---------+--------------------+--------------------+--------------------+--------------------+----------------+
```

```
|   37.857|3 miles east of G...|            -120.086|               False|            Rim Fire|             100|
|34.585595|Angeles National ...|         -118.423176|               False|      Powerhouse Fire|             100|
|     null|                null|                null|                null|                null|            null|
|  33.7095|Hwy 243 & Hwy 74 ...|          -116.72885|               False|       Mountain Fire|             100|
|    39.12|Deadwood Ridge, n...|             -120.65|               False|       American Fire|             100|
|     null|                null|                null|                null|                null|            null|
|     null|                null|                null|                null|                null|            null|
|     null|                null|                null|                null|                null|            null|
|   37.279|Seven miles north...|            -119.318|               False|          Aspen Fire|             100|
|     null|                null|                null|                null|                null|            null|
|     null|                null|                null|                null|                null|            null|
|     2106|                True|The Silver Fire b...|Silver Fire, Augu...|2013-08-07T14:05:00Z|       Finalized|
|    41.32|North Fork of the...|            -123.176|               False|Salmon River Complex|             100|
|   41.035|Tish Tang Ridge e...|            -123.488|               False|      Corral Complex|             100|
| 40.04263|Near Deer Creek, ...|          -121.85397|                True|           Deer Fire|             100|
|     null|                null|                null|                null|                null|            null|
|     True|The Clover Fire b...|Clover Fire, Sept...|2013-09-09T12:32:00Z|           Finalized|              10|
| 32.95435|off Sunrise Hwy, ...|          -116.47381|                True|        Chariot Fire|             100|
|     null|                null|                null|                null|                null|            null|
|     null|                null|                null|                null|                null|            null|
+---------+--------------------+--------------------+--------------------+--------------------+----------------+
only showing top 20 rows

+-----------------+------+--------------------+--------------------+--------------------+--------------------+
|PersonnelInvolved|Public|   SearchDescription|      SearchKeywords|             Started|              Status|
+-----------------+------+--------------------+--------------------+--------------------+--------------------+
|             null|  True|The Rim Fire was ...|Rim Fire, Stanisl...|2013-08-17T15:25:00Z|           Finalized|
|             null|  True|The Powerhouse Fi...|Powerhouse Fire, ...|                null|                null|
|             null|  null|                null|                null|                null|                null|
|             null|  True|The Mountain Fire...|Mountain Fire, Ju...|2013-07-15T13:43:00Z|           Finalized|
|             null|  True|The American Fire...|American Fire, Au...|2013-08-10T16:30:00Z|           Finalized|
|             null|  null|                null|                null|                null|                null|
|             null|  null|                null|                null|                null|                null|
|             null|  null|                null|                null|                null|                null|
|             null|  True|The Aspen Fire bu...|217  Aspen Fire, ...|2013-07-22T22:15:00Z|           Finalized|
|             null|  null|                null|                null|                null|                null|
|             null|  null|                null|                null|                null|                null|
|                8|    40|                null|                null|c400203b-a7fd-4bd...|2013-08-12T18:00:00Z|
|             null|  True|The Salmon River ...|210 Salmon River ...|2013-07-31T22:00:00Z|           Finalized|
|             null|  True|The Corral Comple...|Corral Complex, A...|2013-08-10T11:40:00Z|           Finalized|
|              898|  True|The Deer Fire bur...|Deer Fire, August...|2013-08-23T14:15:00Z|           Finalized|
|             null|  null|                null|                null|                null|                null|
|              201|  null|                null|92af9783-eda9-418...|2013-09-15T07:30:00Z|                null|
|             2147|  True|Chariot Fire burn...|Chariot Fire, Jul...|2013-07-06T12:55:00Z|           Finalized|
|             null|  null|                null|                null|                null|                null|
|             null|  null|                null|                null|                null|                null|
+-----------------+------+--------------------+--------------------+--------------------+--------------------+
```

```
only showing top 20 rows

+-----------------+-------------------+-------------------+--------------------+--------------------+-------------------+
|StructuresDamaged|StructuresDestroyed|StructuresEvacuated|StructuresThreatened|            UniqueId|            Updated|
+-----------------+-------------------+-------------------+--------------------+--------------------+-------------------+
|             null|               null|               null|                null|5fb18d4d-213f-4d8...|2013-09-06 21:30:00|
|             null|               null|               null|                null|                null|               null|
|             null|               null|               null|                null|                null|               null|
|             null|               null|               null|                null|a3149fec-4d48-427...|2013-07-30 21:00:00|
|             null|               null|               null|                null|8213f5c7-34fa-403...|2013-08-30 11:00:00|
|             null|               null|               null|                null|                null|               null|
|             null|               null|               null|                null|                null|               null|
|             null|               null|               null|                null|                null|               null|
|             null|               null|               null|                null|bee8c339-4f26-4b7...|2013-09-24 23:15:00|
|             null|               null|               null|                null|                null|               null|
|             null|               null|               null|                null|                null|               null|
|               20|               null|               null|                null|                null|               null|
|             null|               null|               null|                null|ba76c009-09c9-497...|2013-08-31 09:45:00|
|             null|               null|               null|                null|f3dcbca8-f8ed-46d...|2013-08-12 15:00:00|
|             null|               null|               null|                null|956dbcf6-db40-4b6...|2013-08-29 19:45:00|
|             null|               null|               null|                null|                null|               null|
|             null|               null|               null|                null|                null|               null|
|                9|                149|               null|                null|ee19b2ec-a96a-473...|2013-07-15 09:15:00|
|             null|               null|               null|                null|                null|               null|
|             null|               null|               null|                null|                null|               null|
+-----------------+-------------------+-------------------+--------------------+--------------------+-------------------+
only showing top 20 rows

+------------+
|WaterTenders|
+------------+
|        null|
|        null|
|        null|
|        null|
|        null|
|        null|
|        null|
|        null|
|        null|
|        null|
|        null|
|        null|
|        null|
|        null|
|           8|
|        null|
```

```
|        null|
|          24|
|        null|
|        null|
+------------+
only showing top 20 rows
```

F7 point: check the type of the dataframe

```
type(df)
```

```
pyspark.sql.dataframe.DataFrame
```

F8 point: change the data type and print it

```
#Convert string columns to integer columns
columns_to_convert = ["AcresBurned","AirTankers",
"CountyIds", "CrewsInvolved","Dozers", "Engines", "Helicopters", "Injuries",
"Latitude", "Longitude", "PercentContained",
"PersonnelInvolved","StructuresDamaged", "StructuresDestroyed",
"StructuresEvacuated","StructuresThreatened", "WaterTenders", "Fatalities"
for col_name in columns_to_convert:
    df = df.withColumn(col_name, col(col_name).cast("float"))
#Convert string columns to boolean columns
columns_to_convert = ["Active", "CalFireIncident", "MajorIncident",
'Featured', 'Final', 'Public']
for col_name in columns_to_convert:
    df = df.withColumn(col_name, col(col_name).cast("boolean"))
#Convert string columns to timestamps columns
columns_to_convert = ["Extinguished", "Started", "Updated"]
for col_name in columns_to_convert:
    df = df.withColumn(col_name, col(col_name).cast("timestamp"))
#Convert string column to date
data_for_ml = df.withColumn("ArchiveYear", to_date(col("ArchiveYear"),
"yyyy"))
```

```
root
 |-- AcresBurned: float (nullable = true)
 |-- Active: boolean (nullable = true)
 |-- AdminUnit: string (nullable = true)
 |-- AirTankers: string (nullable = true)
 |-- ArchiveYear: date (nullable = true)
 |-- CalFireIncident: boolean (nullable = true)
 |-- CanonicalUrl: string (nullable = true)
 |-- ConditionStatement: string (nullable = true)
 |-- ControlStatement: string (nullable = true)
 |-- Counties: string (nullable = true)
 |-- CountyIds: float (nullable = true)
 |-- CrewsInvolved: float (nullable = true)
 |-- Dozers: float (nullable = true)
 |-- Engines: float (nullable = true)
 |-- Extinguished: timestamp (nullable = true)
 |-- Fatalities: float (nullable = true)
 |-- Featured: boolean (nullable = true)
 |-- Final: boolean (nullable = true)
 |-- FuelType: string (nullable = true)
 |-- Helicopters: float (nullable = true)
 |-- Injuries: float (nullable = true)
 |-- Latitude: float (nullable = true)
 |-- Location: string (nullable = true)
 |-- Longitude: float (nullable = true)
 |-- MajorIncident: boolean (nullable = true)
 |-- Name: string (nullable = true)
```

```
 |-- PercentContained: float (nullable = true)
 |-- PersonnelInvolved: float (nullable = true)
 |-- Public: boolean (nullable = true)
 |-- SearchDescription: string (nullable = true)
 |-- SearchKeywords: string (nullable = true)
 |-- Started: timestamp (nullable = true)
 |-- Status: string (nullable = true)
 |-- StructuresDamaged: float (nullable = true)
 |-- StructuresDestroyed: float (nullable = true)
 |-- StructuresEvacuated: float (nullable = true)
 |-- StructuresThreatened: float (nullable = true)
 |-- UniqueId: string (nullable = true)
 |-- Updated: timestamp (nullable = true)
 |-- WaterTenders: float (nullable = true)
```

F9 point: Count and Removal of NaN values and duplicates

```
count_Null = [count(when(col(c).isNull(), c)).alias(c) for c in
data_for_ml.columns]
data_for_ml_ null_counts = data_for_ml.select(*count_Null)
data_for_ml = data_for_ml.fillna(0)
data_for_ml.count()
data_for_ml = data_for_ml.dropDuplicates()
data_for_ml.count()
```

F10 point: Removal of irrelevant data

```
data_for_ml = data_for_ml.filter(~col('Started').contains('1969'))
data_for_ml.count()
data_for_ml = data_for_ml.filter(col("Latitude").isNotNull() &
col("Longitude").isNotNull() & col("AcresBurned").isNotNull())
data_for_ml.count()
```

```
1458
1458
```

F11 point: Summary statistics

```
describe_data_for_ml =data_for_ml.describe()
describe_data_for_ml.select(['summary', 'AcresBurned', 'Latitude',
'Longitude']).show()
```

```
+-------+------------------+------------------+-------------------+
|summary|       AcresBurned|          Latitude|          Longitude|
+-------+------------------+------------------+-------------------+
|  count|              1458|              1462|               1461|
|   mean| 1981.153635116598|37.670126051322214|-108.25074648220681|
| stddev|11090.721643309758|143.18026811505163|  36.72815867459589|
|    min|               0.0|          -120.258|         -124.19629|
|    max|          257314.0|            5487.0|           118.9082|
+-------+------------------+------------------+-------------------+
```

F12 point: General Observations

```
data_for_ml.agg({'AcresBurned': 'sum'}).show()
data_for_ml.agg({'PersonnelInvolved': 'sum'}).show()
data_for_ml.agg({'Injuries': 'sum'}).show()
```

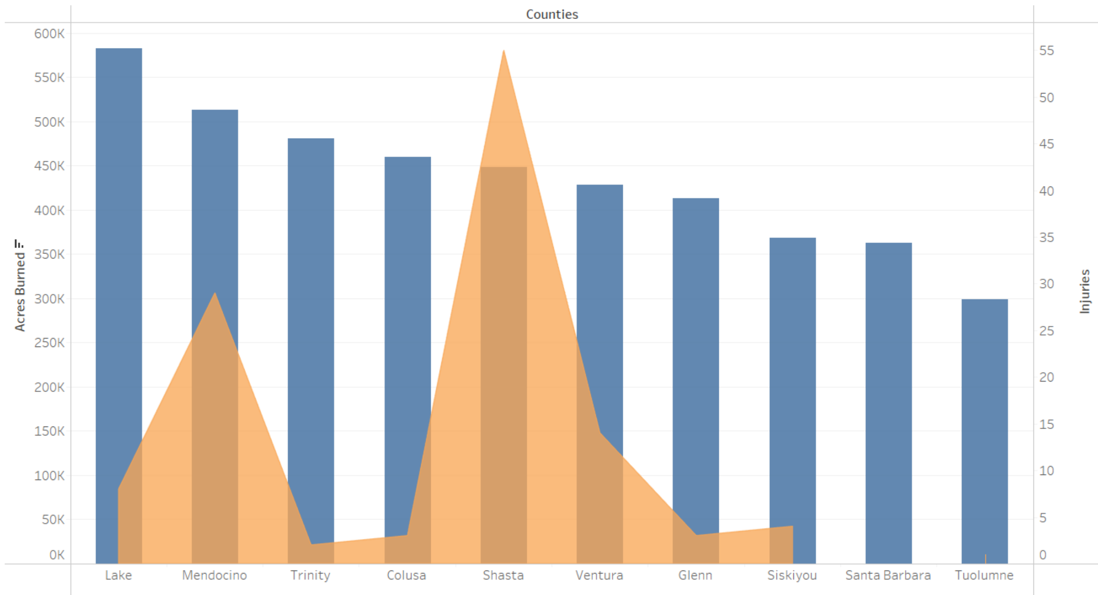
```
+----------------+      +----------------------+      +-------------+
|sum(AcresBurned)|      |sum(PersonnelInvolved)|      |sum(Injuries)|
+----------------+      +----------------------+      +-------------+
|         4961377|      |                 40185|      |          211|
+----------------+      +----------------------+      +-------------+
```

General Observations

Year of Start..	Acres Burned	Injuries	Structures Damaged	Structures Destroyed	Structures Threatened
2013	527,745	113	34	428	176
2014	448,715	137	10	634	2,390
2015	574,503	52	25	381	54
2016	505,927	25	47	774	5,200
2017	1,793,903	31	3,408	17,961	545
2018	3,358,004	28	828	26,855	7,285
2019	285,708	34	202	530	34

F13 point: Top 10 affected counties by Acres Burned and Injuries

Acres Burned and Injuries per County

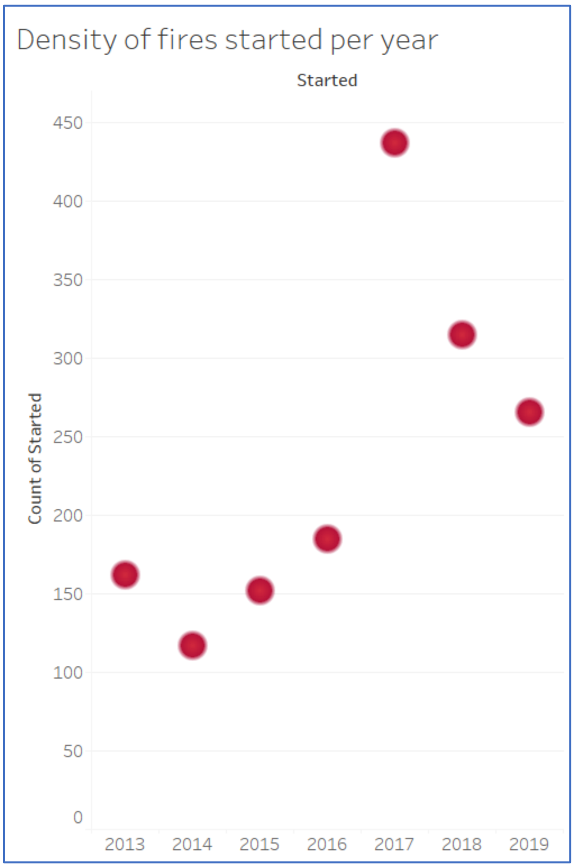


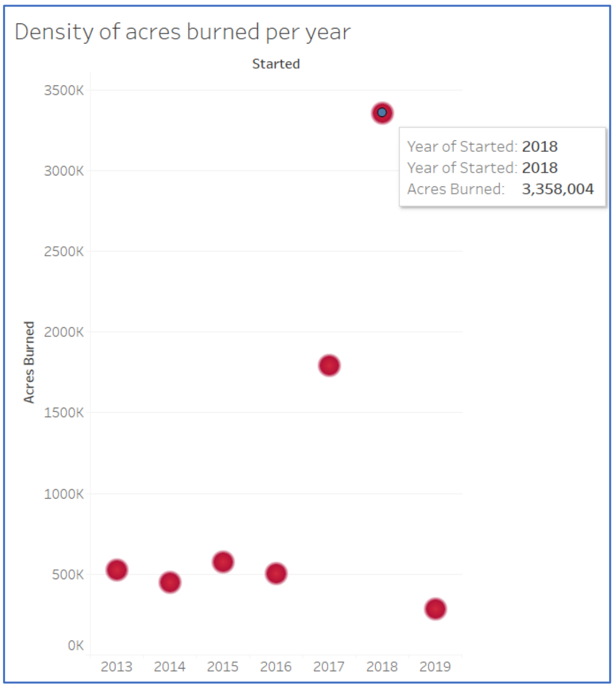
F14 point: Top 10 affected counties by number of fires

```
counties =final_data.groupBy("Counties").count()
counties.orderBy(col("count").desc())
```

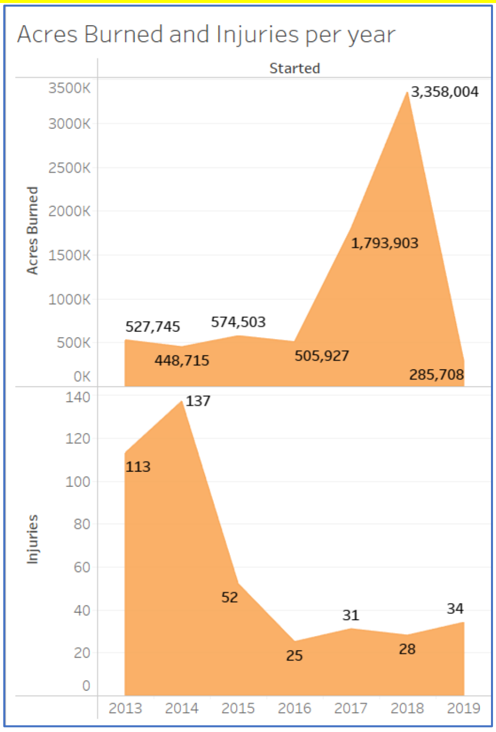
Counties	count
Riverside	140
San Diego	83
San Luis Obispo	61
Kern	59
Shasta	57
Butte	56
Fresno	54
Siskiyou	53
Tehama	49
San Bernardino	49

F15 point: Density of fires started per year

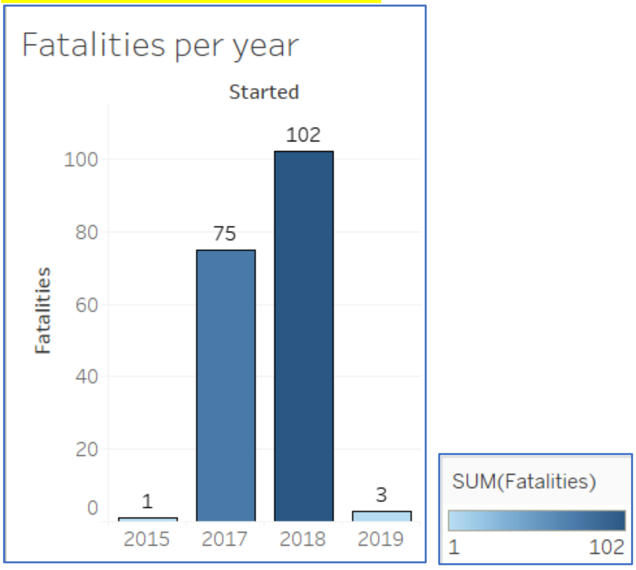




F16 point: Acres burned and injuries per year



F17 point: Fatalities per year



F18 point: Frequency of fires per month

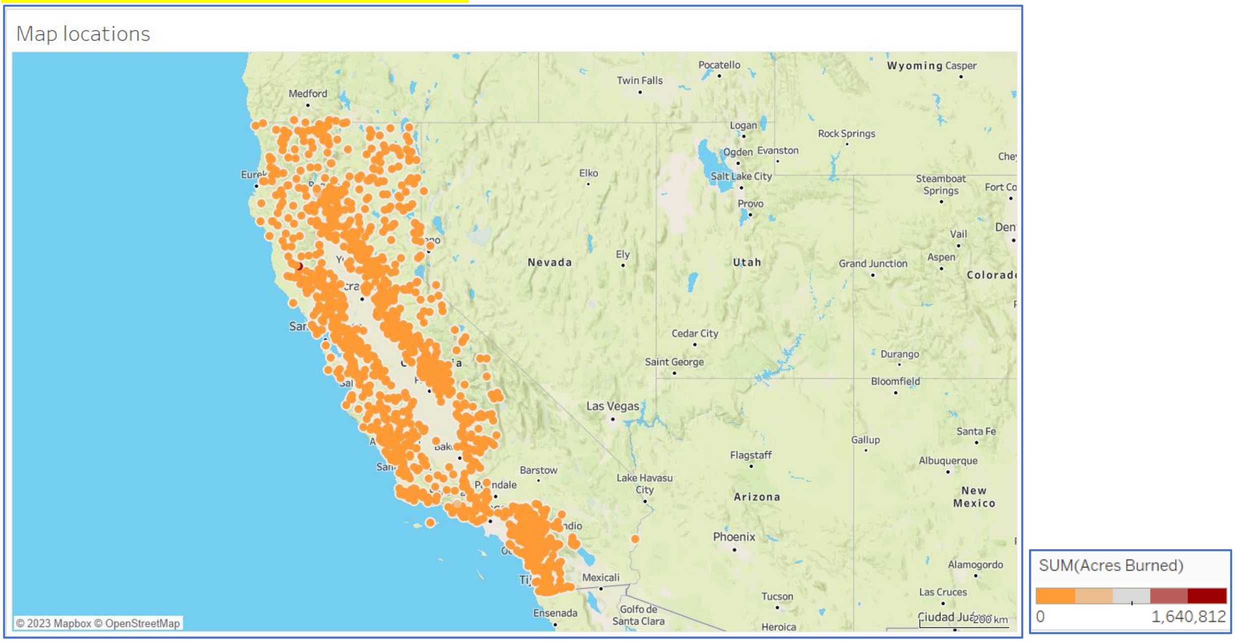
Frequency of fires per month

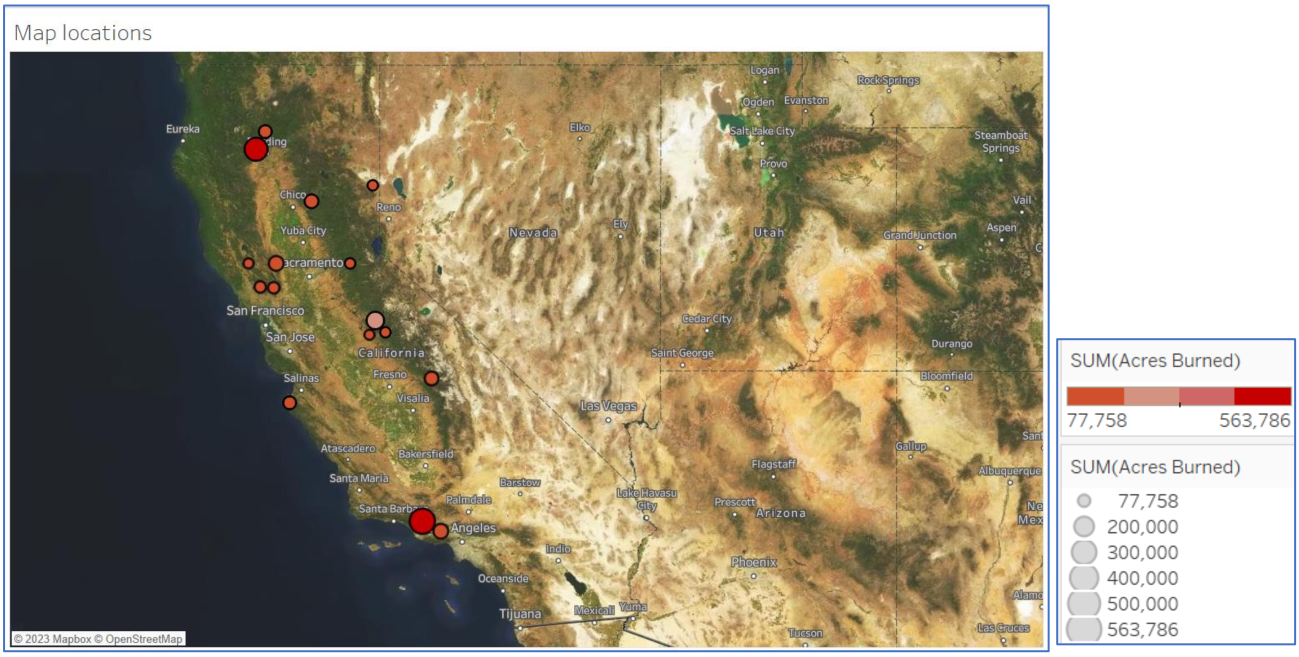
Started

January	Februa..	March	April	May	June	July	August	Septe..	October	Novem..	Decem..
11	10	6	35	140	319	415	282	208	147	42	19

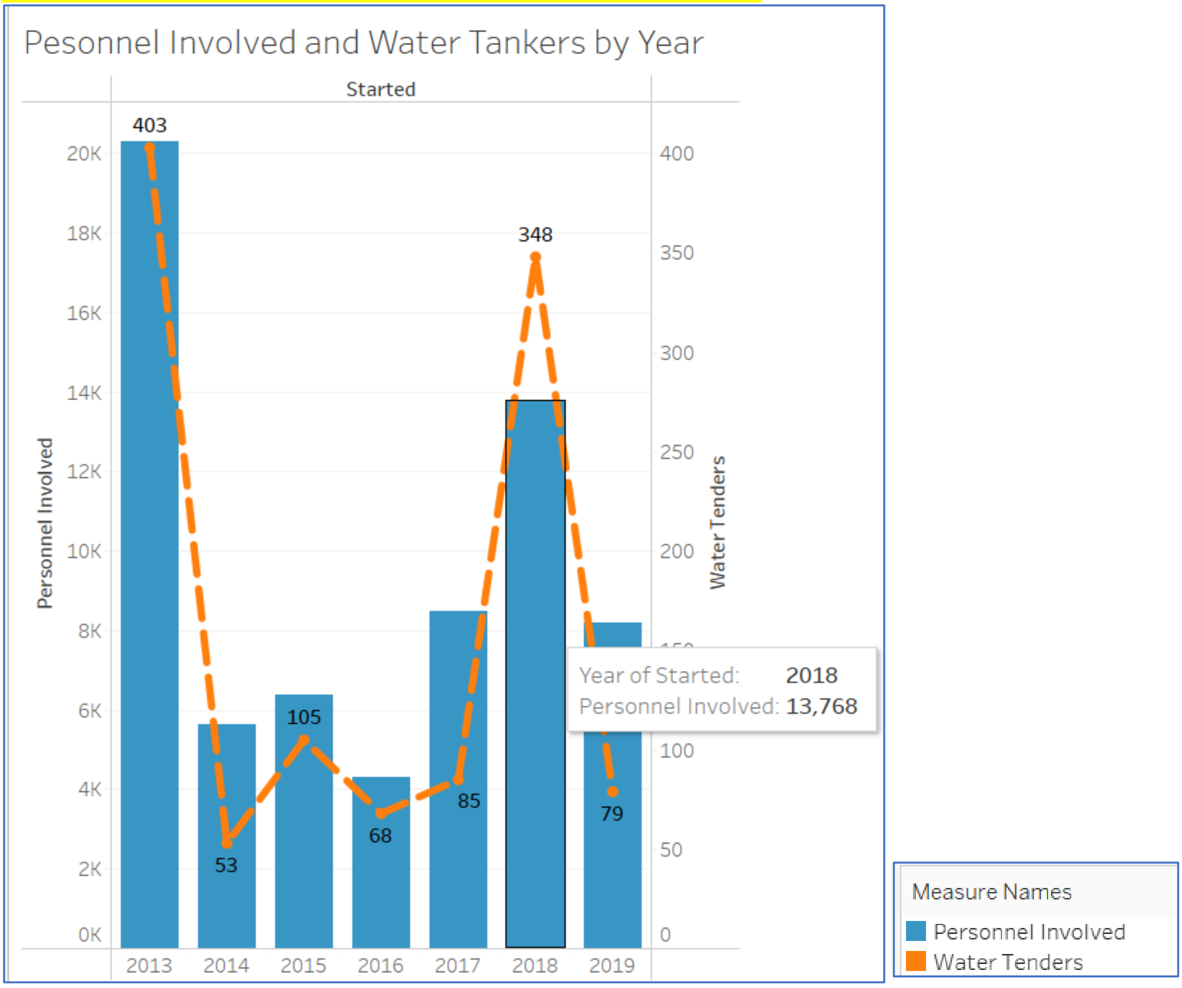
CNT(Started)
6 415

F19 point: Geographical Heat Maps

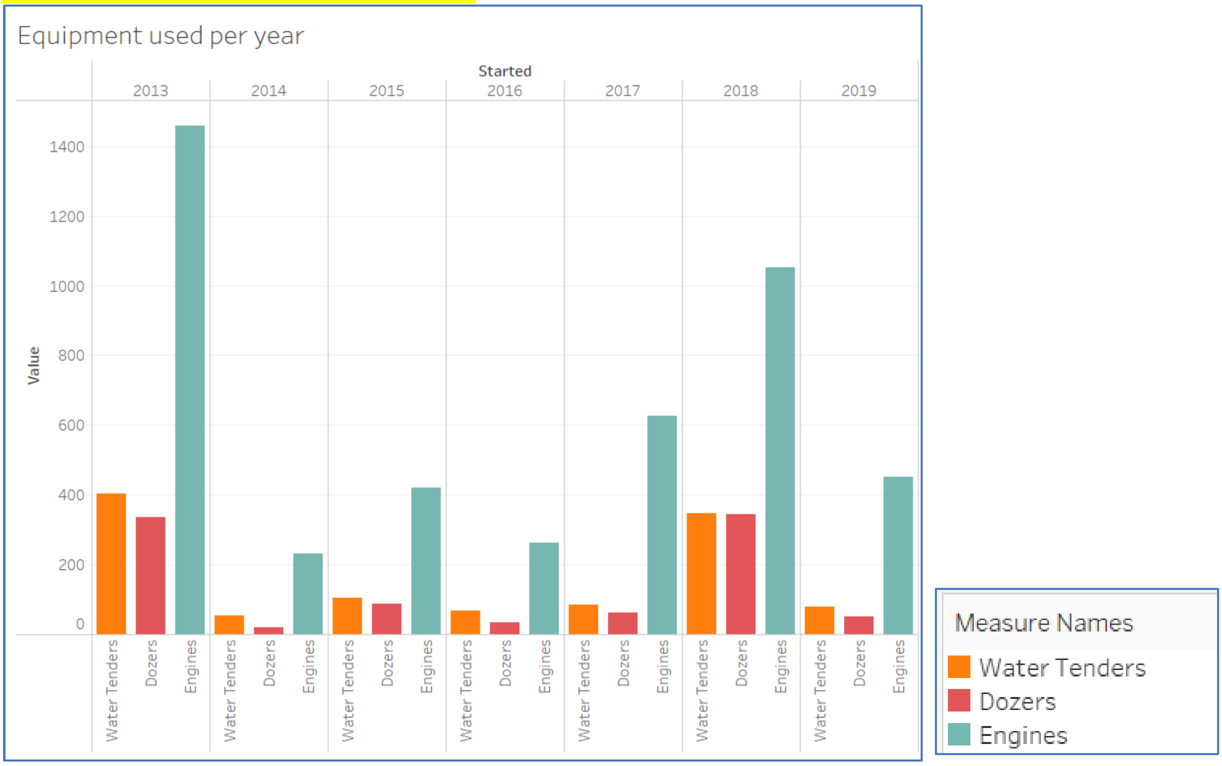




F20 point: Personnel Involved and Water Tankers by year



F21 point: Equipment used per year



F22 point: Ridge Regression

```
from pyspark.ml.regression import LinearRegression
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.evaluation import RegressionEvaluator

data_for_ml = data_for_ml.withColumn("Started",
col("Started").cast("float"))
selected_features = ["Latitude", "Longitude", "Started", "AirTankers",
"WaterTenders", "Helicopters", "PersonnelInvolved"]
assembler = VectorAssembler(inputCols=selected_features,
outputCol="features")
model_data = assembler.transform(data_for_ml)

train_data, test_data = model_data.randomSplit([0.8, 0.2], seed=42)

regressor = LinearRegression(featuresCol="features", labelCol="AcresBurned",
elasticNetParam=0.0)  # Setting elasticNetParam to 0.0 for pure Ridge
regressor_model = regressor.fit(train_data)

predictions = regressor_model.transform(test_data)
predictions.select("Latitude", "Longitude", "AcresBurned",
"prediction").show()

evaluator = RegressionEvaluator(labelCol="AcresBurned",
predictionCol="prediction", metricName="rmse")
rmse = evaluator.evaluate(predictions)
print("Root Mean Squared Error (RMSE):", rmse)
```

```
+----------+----------+-----------+-------------------+
| Latitude| Longitude|AcresBurned|         prediction|
+----------+----------+-----------+-------------------+
| 39.71217|-121.77385|        0.0|1296.3617127643338|
| 35.85253|-120.80411|        0.0|  1618.43369945224|
| 41.86609|-122.73031|        0.0|1601.0741105508478|
|   41.338|  -120.304|        0.0|1576.6011313864146|
| 33.99647|-116.84137|        0.0| 2437.120732161544|
| 39.40972|-121.00056|        2.0|1169.0037931677161|
| 33.90965| -116.9669|       10.0|1579.4529614641524|
| 39.50475|-121.33785|       10.0|1589.0467987910924|
| 35.05747|-120.39295|       10.0|1383.7752280302248|
| 37.12857|-122.12036|       10.0|1306.8677363182069|
|37.178352|-122.07743|       10.0|1136.5687015715102|
|41.632374|-122.37985|       10.0| 1378.302701259825|
| 38.27229|-122.26389|       10.0| 1365.140336444054|
|117.13874| 33.939426|       10.0| 697.0209683551184|
|35.564125|-118.79652|       10.0|1597.0702854044994|
|     37.0|      18.4|       10.0|1403.9315265181885|
| 38.70289|-122.90217|       11.0|  1586.86743736262|
| 38.67474|-121.06088|       13.0| 1587.665188667861|
| 35.80375|-120.52508|       14.0|1604.5110930115952|
| 33.97155|-117.44148|       14.0|1600.1350445283788|
+---------+----------+-----------+------------------+
only showing top 20 rows

Root Mean Squared Error (RMSE): 8967.552848686488
```

F23 point: Ridge Regression

```
from pyspark.ml.regression import LinearRegression
selected_features = ["Latitude", "Longitude", "Started", "AirTankers",
"WaterTenders", "Helicopters", "PersonnelInvolved"]
assembler = VectorAssembler(inputCols=selected_features,
outputCol="features")
model_data = assembler.transform(data_for_ml)
train_data, test_data = model_data.randomSplit([0.8, 0.2], seed=42)
regressor = RandomForestRegressor(featuresCol="features",
labelCol="AcresBurned", numTrees=100)
regressor_model = regressor.fit(train_data)
predictions = regressor_model.transform(test_data)
predictions.select("Longitude", "Latitude", "Injuries", "AcresBurned",
"prediction").show()
evaluator = RegressionEvaluator(labelCol="AcresBurned",
predictionCol="prediction", metricName="rmse")
rmse = evaluator.evaluate(predictions)
print("Root Mean Squared Error (RMSE):", rmse)
```

```
+----------+---------+--------+-----------+------------------+
| Longitude| Latitude|Injuries|AcresBurned|        prediction|
+----------+---------+--------+-----------+------------------+
|-121.77385| 39.71217|     0.0|        0.0| 731.6068832856164|
|-120.80411| 35.85253|     0.0|        0.0| 819.4159439337549|
|-122.73031| 41.86609|     0.0|        0.0|2180.6262557219356|
|  -120.304|   41.338|     0.0|        0.0| 2679.131378463557|
|-116.84137| 33.99647|     0.0|        0.0|1431.9479174996939|
|-121.00056| 39.40972|     0.0|        2.0| 698.9568796667403|
| -116.9669| 33.90965|     0.0|       10.0| 1187.327114543621|
|-121.33785| 39.50475|     0.0|       10.0| 989.7707519734217|
|-120.39295| 35.05747|     0.0|       10.0|1099.3937416549927|
|-122.12036| 37.12857|     0.0|       10.0|505.58666374274827|
|-122.07743|37.178352|     0.0|       10.0| 505.7021113577355|
|-122.37985|41.632374|     0.0|       10.0| 3586.061792139633|
|-122.26389| 38.27229|     0.0|       10.0|1271.3557768339845|
| 33.939426|117.13874|     0.0|       10.0|  5313.02907416288|
|-118.79652|35.564125|     0.0|       10.0|1489.4541182938278|
```

```
|      18.4|     37.0|     0.0|       10.0|1920.4191394266527|
|-122.90217| 38.70289|     0.0|       11.0|1609.8882414246289|
|-121.06088| 38.67474|     0.0|       13.0| 978.9101456450138|
|-120.52508| 35.80375|     0.0|       14.0|1287.9566169412237|
|-117.44148| 33.97155|     0.0|       14.0|  1220.13690465104|
+----------+---------+--------+-----------+------------------+
only showing top 20 rows

Root Mean Squared Error (RMSE): 8757.530616400194
```

F24 point: Count of TRUE and FALSE values

Count of True and False values

Major Incident	
False	1,253
True	383

F25a point: Logistic Regression

```python
from pyspark.sql import SparkSession
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.classification import LogisticRegression
from pyspark.ml import Pipeline
from pyspark.sql.functions import when

feature_columns = ["AcresBurned", "Injuries", "PersonnelInvolved"]
target_column = "MajorIncident"
data = data_for_ml.withColumn(target_column, when(data_for_ml[target_column]
== "TRUE", 1).otherwise(0))
assembler = VectorAssembler(inputCols=feature_columns, outputCol="features")
data = assembler.transform(data)
train_data, test_data = data.randomSplit([0.7, 0.3], seed=123)
lr = LogisticRegression(featuresCol="features", labelCol=target_column)
pipeline = Pipeline(stages=[lr])
model = pipeline.fit(train_data)
predictions = model.transform(test_data)
predictions.select("features", target_column, "probability",
"prediction").show(predictions.count(), truncate=False)
```

```
+--------------+-------------+--------------------+----------+
|      features|MajorIncident|         probability|prediction|
+--------------+-------------+--------------------+----------+
|     (3,[],[])|            1|[0.89614127907365...|       0.0|
|     (3,[],[])|            0|[0.89614127907365...|       0.0|
|     (3,[],[])|            0|[0.89614127907365...|       0.0|
|     (3,[],[])|            0|[0.89614127907365...|       0.0|
|     (3,[],[])|            0|[0.89614127907365...|       0.0|
|     (3,[],[])|            0|[0.89614127907365...|       0.0|
| [2.0,0.0,0.0]|            0|[0.89615421922678...|       0.0|
|[10.0,0.0,0.0]|            0|[0.89620596558587...|       0.0|
|[10.0,0.0,0.0]|            0|[0.89620596558587...|       0.0|
|[10.0,0.0,0.0]|            0|[0.89620596558587...|       0.0|
|[10.0,0.0,0.0]|            0|[0.89620596558587...|       0.0|
|[10.0,0.0,0.0]|            1|[0.89620596558587...|       0.0|
|[10.0,0.0,0.0]|            0|[0.89620596558587...|       0.0|
|[10.0,0.0,0.0]|            0|[0.89620596558587...|       0.0|
|[10.0,0.0,0.0]|            0|[0.89620596558587...|       0.0|
|[10.0,0.0,0.0]|            0|[0.89620596558587...|       0.0|
|[10.0,0.0,0.0]|            0|[0.89620596558587...|       0.0|
|[10.0,0.0,0.0]|            0|[0.89620596558587...|       0.0|
|[10.0,0.0,0.0]|            0|[0.89620596558587...|       0.0|
|[10.0,0.0,0.0]|            0|[0.89620596558587...|       0.0|
```

```
+--------------+-------------+--------------------+----------+
only showing top 20 rows
```

F25b point: Logistic Regression Evaluation

```
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
evaluator = MulticlassClassificationEvaluator(labelCol=target_column)
precision = evaluator.evaluate(predictions, {evaluator.metricName:
"weightedPrecision"})
recall = evaluator.evaluate(predictions, {evaluator.metricName:
"weightedRecall"})
f1_score = evaluator.evaluate(predictions, {evaluator.metricName: "f1"})
valuator_rmse = RegressionEvaluator(labelCol=target_column,
predictionCol="prediction", metricName="rmse")
rmse = evaluator_rmse.evaluate(predictions)
print("Classification Report:")
print(f"Weighted Precision                    : {precision:.4f}")
print(f"Weighted Recall                       : {recall:.4f}")
print(f"F1-Score                              : {f1_score:.4f}")
```

```
Classification Report:
Weighted Precision              : 0.8916
Weighted Recall                 : 0.8797
F1-Score                        : 0.8594
```

References

[1] Adolphe, J. (2018). *Why are California wildfires so bad? An interactive look*. [online] the Guardian. Available at: https://www.theguardian.com/world/ng-interactive/2018/sep/20/why-are-california-wildfires-so-bad-interactive.

[2] Beheshti, N. (2022). *Random Forest Regression*. [online] Medium. Available at: https://towardsdatascience.com/random-forest-regression-5f605132d19d.

[3] CalFire (n.d.). *About | CAL FIRE*. [online] www.fire.ca.gov. Available at: https://www.fire.ca.gov/about [Accessed 18 Aug. 2023].

[4] CHRIS X (n.d.). *Wildfires - Geospatial Visualization and EDA*. [online] kaggle.com. Available at: https://www.kaggle.com/code/docxian/wildfires-geospatial-visualization-and-eda [Accessed 19 Aug. 2023].

[5] EJHUERTA (n.d.). *California Wildfire Analysis Years 2013-2019*. [online] kaggle.com. Available at: https://www.kaggle.com/code/ejhuerta/california-wildfire-analysis-years-2013-2019 [Accessed 19 Aug. 2023].

[6] freeCodeCamp.org (2021). *PySpark Tutorial*. [online] www.youtube.com. Available at: https://www.youtube.com/watch?v=_C8kWso4ne4&pp=ygURcHlzcGFyayB0dXRvcmlhbCA%3D [Accessed 19 Aug. 2023].

[7] https://montgomerycountypolicereporter.com/author/scottjengle (2020). *MONTGOMERY COUNTY FIREFIGHTERS DEPART FOR CALIFORNIA WILDFIRES*. [online] montgomery county police reporter. Available at: https://montgomerycountypolicereporter.com/montgomery-county-firefighters-depart-for-california-wildfires/ [Accessed 18 Aug. 2023].

[8] IBM (n.d.). *What is Logistic regression? | IBM*. [online] www.ibm.com. Available at: https://www.ibm.com/topics/logistic-regression#:~:text=Resources-.

[9] LESLEYDING (n.d.). *California Forest Fire Spatial Analysis*. [online] kaggle.com. Available at: https://www.kaggle.com/code/lesleyding/california-forest-fire-spatial-analysis [Accessed 19 Aug. 2023].

[10] Naveen (NNK) (2021). *PySpark - Find Count of null, None, NaN Values*. [online] Spark by {Examples}. Available at: https://sparkbyexamples.com/pyspark/pyspark-find-count-of-null-none-nan-values/.

[11] Team, G.L. (2020). *Ridge Regression Definition & Examples | What is Ridge Regression?* [online] GreatLearning Blog: Free Resources what Matters to shape your Career! Available at: https://www.mygreatlearning.com/blog/what-is-ridge-regression/.