National College of Ireland

Project Submission Sheet

Student Name:	Kiruthika Suresh		
Student ID:	23188916		
Programme:	Master's in data Analytics	**Year:**	January 2024
Module:	Data Intensive Architectures		
Lecturer:	Jaswinder Singh		
Submission Due Date:	18-05-2024		
Project Title:	Project Report		
Word Count:	3368		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	Kiruthika Suresh
Date:	20-05-2024

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

I. [INSERT MODULE NAME]

II. [INSERT TITLE OF YOUR ASSIGNMENT]

Your Name/Student Number	Course	Date

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

III. AI ACKNOWLEDGMENT

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool

DESCRIPTION OF AI USAGE

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

[Insert Tool Name]	
[Insert Description of use]	
[Insert Sample prompt]	[Insert Sample response]

EVIDENCE OF AI USAGE

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

ADDITIONAL EVIDENCE:

[Place evidence here]

ADDITIONAL EVIDENCE:

COMPARATIVE ANALYSIS OF THE ROAD ACCIDENT DATASET

Kiruthika Suresh
Student Id - 23189916
Data Analytics, School of Computing
National College of Ireland
x23189916@student.ncirl.ie

***Abstract*—In this Project we undertake a comparative analysis of the road accident through the application of big data analysis. We have a large dataset with about 1.6 million accident records from the year (2009 - 2014) happened in United Kingdom which consists of various information's like severity, number casualties, weather conditions, road types, etc., with these two datasets by using Apache Spark, mainly on Pyspark, using map reduce programming model we perform various Exploratory analysis by using RDD (Resilient Distributed Dataset) which when datasets like these are stored it will process it and analyze the data by grouping them in a meaningful cluster. Spark is an independent cluster management framework supporting Python, Java, Scala, and R, in which we are going to use PySpark with various libraries for our big data analysis. By analyzing this dataset, we are going to find various insights on road accidents which will help us in future road safety management. We will find the objective question and provide a quality idea for the wellness of road Safety.**

Keywords—Big Data, PySpark, MapReduce, RDD, Road Accident.

A. INTRODUCTION

In this modern world, the digital platforms have grown so far with this evolution of social media, cloud computing, mobile phones, laptops large amount of data are created every minute. With this growing data we could not able to manage this amount of billions and billions of data. So, to manage this large amount of data we introduced Data-intensive architectures which consist of various technologies and frameworks designed. The big data can be handled with these technologies and frameworks not only that we can also analyse real time data and provide machine learning and artificial intelligence solutions. It not only handles large data but also provides ideas for us to make decisions and innovations.

A. Project Objectives:

In this study we evaluate the road accidents of two time periods (2009 to 2014) with the goal to identify the temporal patterns and the factors that contributing to these accidents. As we see in our day-to-day life that there is an drastic increase in road accidents it is important to analyze the factors and implement preventive measures. The specific aim of our projects is:

1. Identifying the temporal trends and variations:

   - By comparing the two road accidents dataset of year 2009 to 2014 we can identify the trends and patterns of the datasets over time.
   - Identifying the variations and trends in accident rates, severity and other factors such as light conditions, road types and weather conditions.

2. Understanding the impact of Police Attendance in the scene:

   - We are going to investigate that whether the police attendance at the accident scene is impacting the accident reporting or not
   - By analyzing this we are going to find the completeness of the accident data with or without the presence of police.

3. Identifying the main key factors influencing road accidents:

   - We are going to identify the factors that are causing accidents to happen in the first place and how they are evolving over the years.
   - Providing an idea of how these factors are contributing to accidents so that we can provide insights into the road safety measures and suggest preventive measures.

B. Scope and Challenges:

1. Scope:

The main aim of this project is to provide insights on the road safety by doing a comparative analysis on the road accident dataset of over 1.6 million accidents happened in various years detailing aspects like severity, number of casualties, weather conditions, road types and many more.

2. Challenges:

The challenges that we face in this project are:

- **Data Quality:**

  We must ensure that the data is clean and consistent for that we must handle missing values, null values, outliers, and we must transform our dataset in a way that we could achieve our objective questions.

- **Data Volume:**

  We must handle this large amount of data which is very challenging. we should be able to manage and process this large data efficiently using Map reduce where it will process and analyses our data into logical clusters.

C. Innovation and Importance:

In this project we use Apache Spark for Distributed data processing to handle the wide range of datasets. Where using PySpark's data frame and RDD's ensures comprehensive and scalable data analysis. We use various machine learning algorithms and statistical techniques to find the patterns and correlations in the dataset. We use libraries like Seaborn and Matplotlib to visualize our data, to present our findings in a neat manner and to make it understandable for everyone.

The importance of this project is to contribute our findings to enhance road safety and give valuable ideas that we gained from our analysis helping the environment for better life. By solving this complex data, we show that using data intensive architectures we can handle big complex data in solving real-world problems. We provide deep understanding on identifying the various factors that influence the road accidents.

D. Structure of the Paper:

This report will flow in this structure where we give introduction, Data Source, Methodology, Implementation and Architecture, Results, Conclusion and Future works and References

B. DATA SOURCE

The dataset that we have taken for our analysis is sourced from the United Kingdom Government, which consist of road accident records from 2009 to 2016. This dataset includes approximately 1.6 million recorded accidents records, making this dataset a complex yet challenging dataset for our analysis. This data is divided into two CSV files namely:

1. **Accidents_2009_to_2011.csv**
2. **Accidents_2012_to_2014.csv**

Each dataset comprises 33 columns, detailing various aspects of each accidents,

1. Accident_Index,
2. Locating_Easting_OSGR,
3. Location_Northing_OSGR,
4. Longitude,
5. Latitude,
6. Police_Force,
7. Accident_Severity,
8. Number_of_Vehicles,
9. Number_of_Casualties,
10. Date,
11. Day_of_Week,
12. Time,
13. Local_Authority_(District),
14. Local_Authority_(Highway),
15. 1st_Road_Class,
16. 1st_Road_Number,
17. Road_Type,
18. Speed_limit,
19. Junction_Detail,
20. Junction_Control,
21. 2nd_Road_Class,
22. 2nd_Road_Number,
23. Pedestrian_Crossing-Human_Control,
24. Pedestrian_Crossing-Physical_Facilities,
25. Light_Conditions,
26. Weather_Conditions,
27. Road_Surface_Conditions,
28. Special_Conditions_at_Site,
29. Carriageway_Hazards,
30. Urban_or_Rural_Area,
31. Did_Police_Officer_Attend_Scene_of_Accident,
32. LSOA_of_Accident_Location,
33. Year.

This dataset is from the Kaggle website https://www.kaggle.com/datasets/daveianhickey/2000-16-traffic-flow-england-scotland-wales. The dataset provide detailed information on road accidents which are essential for analysing temporal trends, the impact of police attendance and identifying key factor influencing accidents.

A. Literature Review:

The paper gives information on Apache Spark, [1] it is an open-source distributed computing framework, revolutionized Industry 3.0 and Industry 4.0 with its agility and unified analytics capabilities. Originating from the University of Berkeley in 2009 and later adopted by the Apache Software Foundation, Spark handles well-structured big datasets with lightning-fast Speed. It offers implicit data parallelism, robust fault tolerance, and ease of maintenance through Application Interfaces APIs). Apache Spark, while leveraging some aspects of Apache Hadoop, operates with its independent cluster management framework. It boasts language-agnostic interfaces, supporting Python, Java, Scala, and R, and offers a suite of tools and components for high-performance parallel distributed computing on clusters. Notably, it finds extensive utilization in data processing among a substantial majority of Fortune 500 tech unicorns. [1].

This paper gives idea on RDD, [2] Compared to Hadoop, Apache Spark operates 100 times faster in memory and 10 times faster on the disk. It is the most effective and user-friendly tool for addressing large data issues. In addition to being designed in Scala, Spark has APIs for R, Python, and Java, among other computer languages. The main emphasis in spark components is still on features and abstraction. The spark primarily features machine learning procedures for planning and implementing. Pipelines and model tuning are examples of machine learning components. Graph analysis and streaming are

also possible to process. Spark is used by large corporations in social networks and commerce to handle large amounts of data. RDD robust distributed datasets are the two primary abstractions offered by Spark Architect. There is a set of data in RDD that is separated into distinct partitions and could be kept in working nodes' memory within clusters. Additionally, Spark parallelizes collection based on Scala and supports several RDD types for files stored on HDFS. Its operations include transformation and action. [2]

C. METHODOLOGY

A. Data Collection and Preprocessing:

For our analysis we use amazon s3, which is a cloud storage system to store our large accident dataset. We create an amazon bucket and load our dataset, and then we use Data Bricks which is a cloud based platform used for analysis of data. It was made by Apache Spark's initial developers. It supports businesses in developing, growing and regulating AI and data. We will first create a new compute through which we will run our Apache Spark code. We have a worker node which is an executor process and the driver process. In which the driver process has 16 GB RAM Memory. In this Analysis we are going to use Python Language that is PySpark for our analysis. First, we are going to create a spark session after which we load our dataset from amazon s3 to data bricks, into a data frame. After which we proceed with Data Preprocessing.

When the data is imported to our data frame the data types of our data are all in string hence, we need to change it as per our requirement. Hence, we create a new schema and give the data type that we need. Then we find the shape of our 2 datasets in which the 1 dataset contains 469442 rows and 33 columns, and 2 dataset contains 464697 rows and 33 columns. We combine our two data frames into one combined data frame using SQL function and now the combined data frame consists of 934139 rows and 33 columns.

```
combined_df:  pyspark.sql.dataframe.DataFrame
    Accident_Index: string
    Location_Easting_OSGR: integer
    Location_Northing_OSGR: integer
    Longitude: float
    Latitude: float
    Police_Force: integer
    Accident_Severity: integer
    Number_of_Vehicles: integer
    Number_of_Casualties: integer
    Date: string
    Day_of_Week: integer
    Time: string
    Local_Authority_(District): integer
    Local_Authority_(Highway): string
    1st_Road_Class: integer
    1st_Road_Number: integer
    Road_Type: string
    Speed_limit: integer
    Junction_Detail: float
    Junction_Control: string
    2nd_Road_Class: integer
    2nd_Road_Number: integer

Shape of the Dataset: (934139, 33)
```

Figure 1

In Figure 1, we could see the combined data frame's output with all the data types and shape of the dataset. As a process of Data Cleaning, Initially we check for all the null values that are present in our dataset. In which Time, Junction_Detail, Junction_Control, Weather_Conditions, Road_Surface_Conditions, Special_Conditions_at_Site, Carriageway_Hazards, Did_Police_Officer_Attend_Scene_of_Accident, LSOA_of_Accident_Location consist of null values as in Figure 2.

```
|Accident_Index|Location_Easting_OSGR|Location_Northing_OSGR|Longitude|Latitude|Police_Force|Accident_Severity|Number_of_Vehicles|Number_of_Casualties|Date|Day_of_
Week|Time|Local_Authority_(District)|Local_Authority_(Highway)|1st_Road_Class|1st_Road_Number|Road_Type|Speed_limit|Junction_Detail|Junction_Control|2nd_Road_Class
|2nd_Road_Number|Pedestrian_Crossing-Human_Control|Pedestrian_Crossing-Physical_Facilities|Light_Conditions|Weather_Conditions|Road_Surface_Conditions|Special_Cond
itions_at_Site|Carriageway_Hazards|Urban_or_Rural_Area|Did_Police_Officer_Attend_Scene_of_Accident|LSOA_of_Accident_Location|Year|
|            0|                  0|                   0|       0|      0|          0|               0|                 0|                  0|   0|   0|
0|  50|                        0|                        0|            0|              0|       0|          0|         934139|          365890|            0|
0|                              0|                                      0|               0|               106|                   1296|                      4|
6|                  0|                                        547|               60727|   0|
```

Figure 2

In order to remove these null values first, I drop the unnecessary columns such as Junction_Detail, and Junction_Control. After which I took mode to impute all the missing values. And then when I check for null values my dataset have 0 null values.

B. Analytical Approch:

In spark a set of our data is partitioned and are stored into the working nodes of cluster. By distributing the computation across multiple nodes, using map reduce process large datasets parallelly, reducing time for data analysis. We convert our combined data frame into RDD. Our input data is divided into chunks which are processed in parallel by different node clusters. After the map phase the key values are shuffled and sorted. After which in reduced phase the grouped data is processed by applying reduce function which gives the results as per our research question.

D. IMPLEMENTATION AND ARCHITECTURE

1. System Design:

The dataset ‘Accident_2009_to_2011.csv’ and ‘Accident_2012_to_2014.csv’ are initially stored in a cloud based storage solution such as amazon s3. We created a bucket and loaded our dataset into amazon S3.

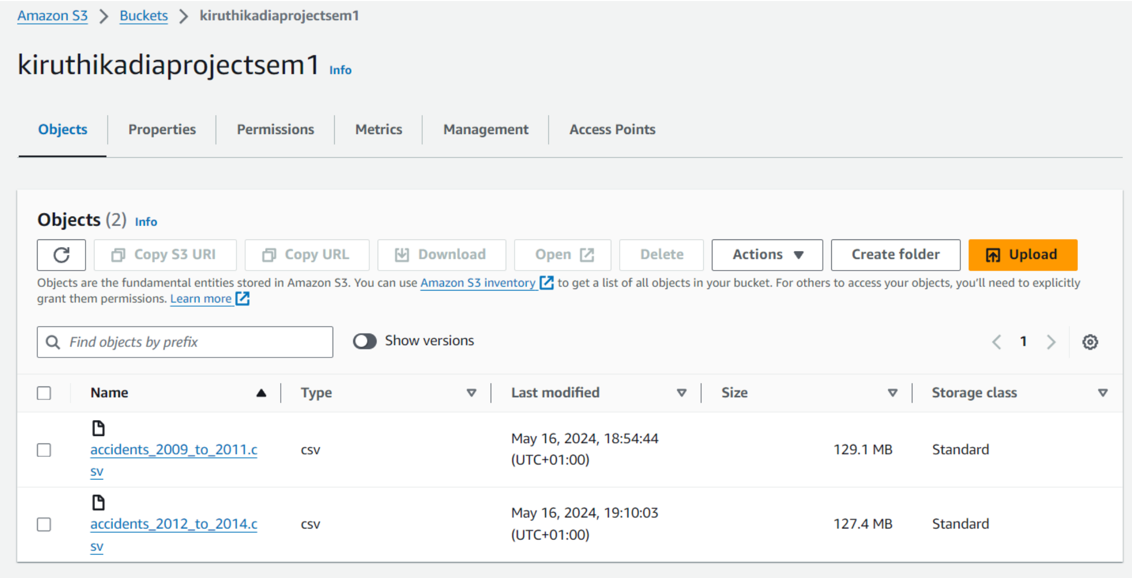

Figure 3

Apache Spark is used as data processing framework. Resilient Distributed Dataset (RDD) enable efficient handling of large dataset. Using map reduce we do parallel data processing tasks, we handled the missing values, null values and outliers in our dataset. Then as a part of transformation we standardize the formats of the date column we convert it into date format and abstract month and year column from it.

```
---+------------------+------------------+-----------------------------------------------+------------------------+----+-----+
|Accident_Index|Location_Easting_OSGR|Location_Northing_OSGR|Longitude|Latitude|Police_Force|Accident_Severity|Number_of_Vehicles|Number_of_Casualties|      Date|D
ay_of_Week| Time|Local_Authority_(District)|Local_Authority_(Highway)|1st_Road_Class|1st_Road_Number|         Road_Type|Speed_limit|2nd_Road_Class|2nd_Road_Number|
Pedestrian_Crossing-Human_Control|Pedestrian_Crossing-Physical_Facilities|    Light_Conditions|  Weather_Conditions|Road_Surface_Conditions|Special_Conditions_at_S
ite|Carriageway_Hazards|Urban_or_Rural_Area|Did_Police_Officer_Attend_Scene_of_Accident|LSOA_of_Accident_Location|Year|Month|
+--------------+---------------------+----------------------+---------+--------+------------+-----------------+------------------+--------------------+----------+-
---------+-----+--------------------------+-------------------------+--------------+---------------+------------------+-----------+--------------+---------------+
---------------------------------+---------------------------------------+--------------------+--------------------+-----------------------+--------------------
---+------------------+------------------+-----------------------------------------------+------------------------+----+-----+
| 200901BS70001|               524910|                180800|-0.201349|51.51227|           1|                2|                 2|                   1|2009-01-01|
5|15:11|                        12|                E09000020|             6|              0|    One way street|         30|             6|              0|
None within 50 me...|                         No physical cross...|Daylight: Street ...|Fine without high...|                   Dry|                   None|
None|                  1|                                          Yes|                E01002882|2009|    1|
| 200901BS70002|               525050|                181040|-0.199248| 51.5144|           1|                2|                 2|                  11|2009-05-01|
2|10:59|                        12|                E09000020|             5|              0|Single carriageway|         30|             5|              0|
None within 50 me...|                               Zebra crossing|Daylight: Street ...|Fine without high...|              Wet/Damp|                   None|
None|                  1|                                          Yes|                E01002886|2009|    5|
+--------------+---------------------+----------------------+---------+--------+------------+-----------------+------------------+--------------------+----------+-
```

Figure 4

From Figure 4, we could see the Date and Year column created at the end. Then we analyze the important columns in our dataset to get valuable information. First, we group the Accident severity and calculate the sum of total number of casualties. From Figure 5, we could see the relationship between them to achieve our research question.

Accident_Severity	Total_Casualties
1	19856
3	1049988
2	185001

Figure 5

And then we also analyzed the relationship by grouping the Light conditions and calculated the total number of Casualties as in figure 6.

Weather_Conditions	Total_Casualties
Raining without h...	150576
Snowing with high...	2012
Snowing without h...	11219
Unknown	20900
Other	27072
Fine with high winds	14779
Fine without high...	1003748
Raining with high...	17604
Fog or mist	6935

Figure 6

We also found the insight by finding the relationship between year column and total casualties from figure 7.

Year	Total_Casualties
2009	222146
2010	208648
2011	203950
2012	241954
2013	183670
2014	194477

Figure 7

Spark's RDD provides fault tolerance, which allows the system to recompute lost data partitions in case of node failures which provides robustness to our project.

2. Tools Selection and Workflow:

To efficiently process and analyze our large-scale road accident dataset we use a combination of open-source tools. Tools are used to ensure scalability, robustness and efficient data handling. These tools are used for our analysis, Apache Spark which is the core of data processing, and it handles data Ingestion, transformation, cleaning and analysis through its RDD. Cloud storage systems such as Amazon S3 is used to load the data into cloud storage system. Data Brick is the platform where we perform Spark for data processing. For visualization we used libraries like Matplotlib and Seaborn.

The workflow of our Data Analysis flows through data ingestion where we load our dataset from amazon s3 to Py Spark. Then we clean the data and preprocess it for transformation in data cleaning we handle missing values, standardize the data formats and combine the datasets. We perform the initial analysis to understand our data distribution and pattern. For comprehensive data analysis we use advanced techniques and machine learning models like Spark MLlib and then we visualize our dataset using tools like Matplotlib and Seaborn.

E. RESULTS

1. Temporal Trends in Road Accidents:

From Figure 8, After analyzing the data from 2009 to 2014 we observe that from 2009 to 2011 there is a gradual decrease in number of the road accidents over the years but in 2012 there is a drastic increase in road accidents and followed by that there is a drastic decrease.

To bring this visualization we used several steps in which first we extracted the year and total number of casualties per year and then we aggregate the total number of casualties per year using RDD then we convert the RDD to a Spark Data frame. Then we convert the data frame to Pandas data frame. Then we plot a line plot for this analysis.

We expected a reduction in accidents due to the advancements in vehicle safety technology and strict traffic regulations. The findings confirm this hypothesis indicating a positive impact but still we could see that the 2nd dataset is not reliable for our analysis as it shows a drastic rise and drastic fall on road accidents.

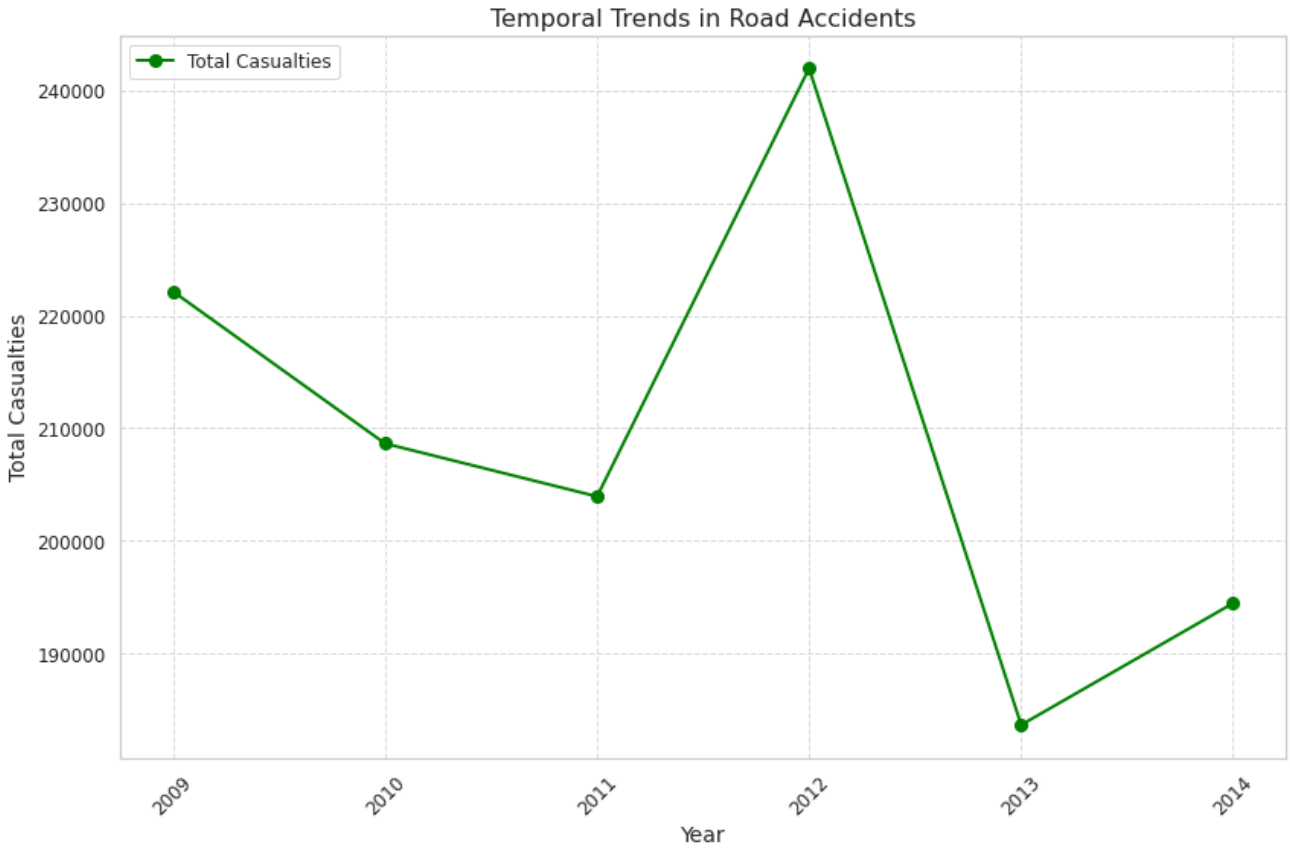

Figure 8

2. Impact of police attendance on Accident Reporting:

From Figure 9, After Analysis, the presence of police officers at the scene of accidents was accurate and complete data reporting. Accidents attended by police had fewer missing values and more detailed records.

To analyze this research question we used various steps, like extracting the necessary data from RDD and then converting the RDD to a Pandas Data Frame and we used a box plot to visualize this data accurately.

It was expected that the police attendance at the accident scene would improve the quality of the data and the findings are also proves that with the expectation, which shows the importance of police presence at the scene of accident.

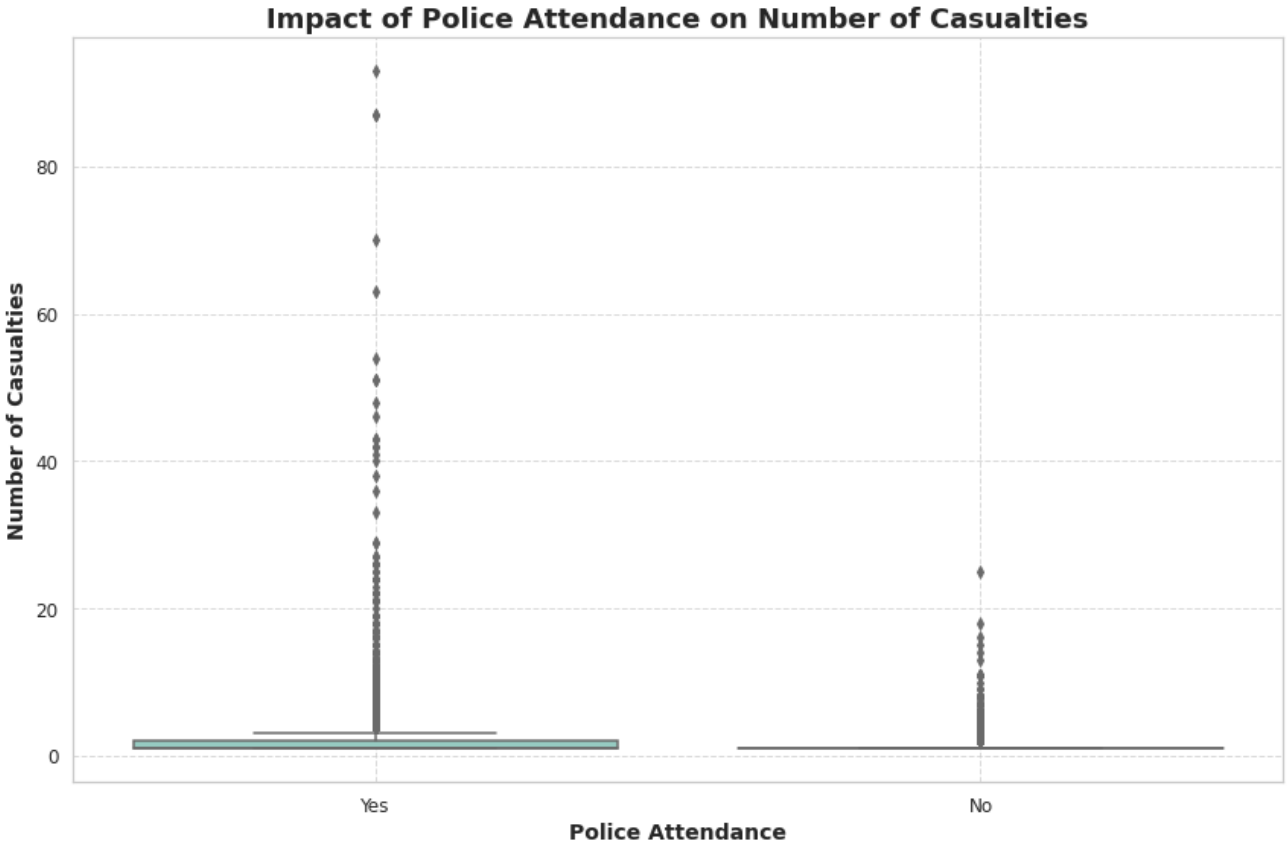

Figure 9

3. *Factors Influencing Road Accidents:*

The key factors that are influencing Road accidents are Speed limit, weather condition, lighting conditions, and road types.

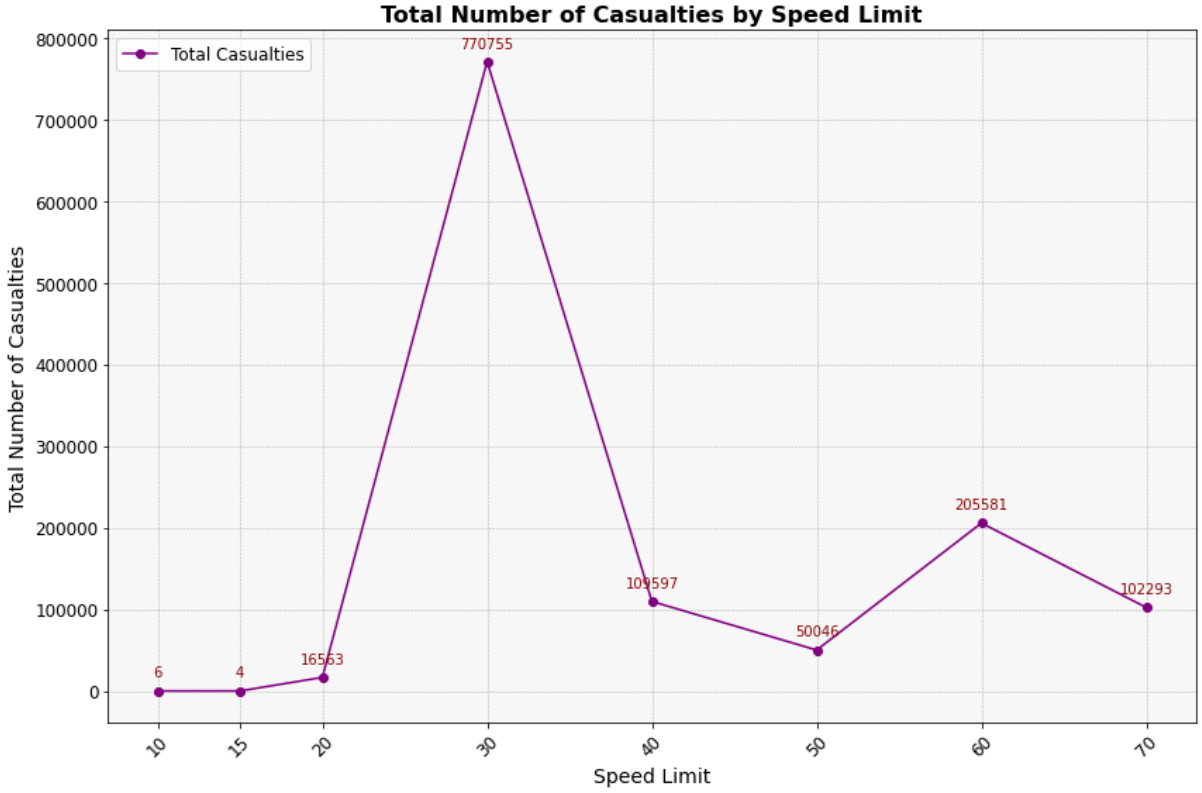

Figure 10

From figure 10, after combining the two datasets with Spark SQL, we are finding the relationship between the speed limit of the vehicle and the number of casualties. We first group the speeding limit and find the sum of the number of casualties. Then convert the PySpark data frame to Pandas Date Frame for visualization. Then we plot the graph using line plot. By this plot we could see that the total number of casualties are more when the speed limit is at 30.

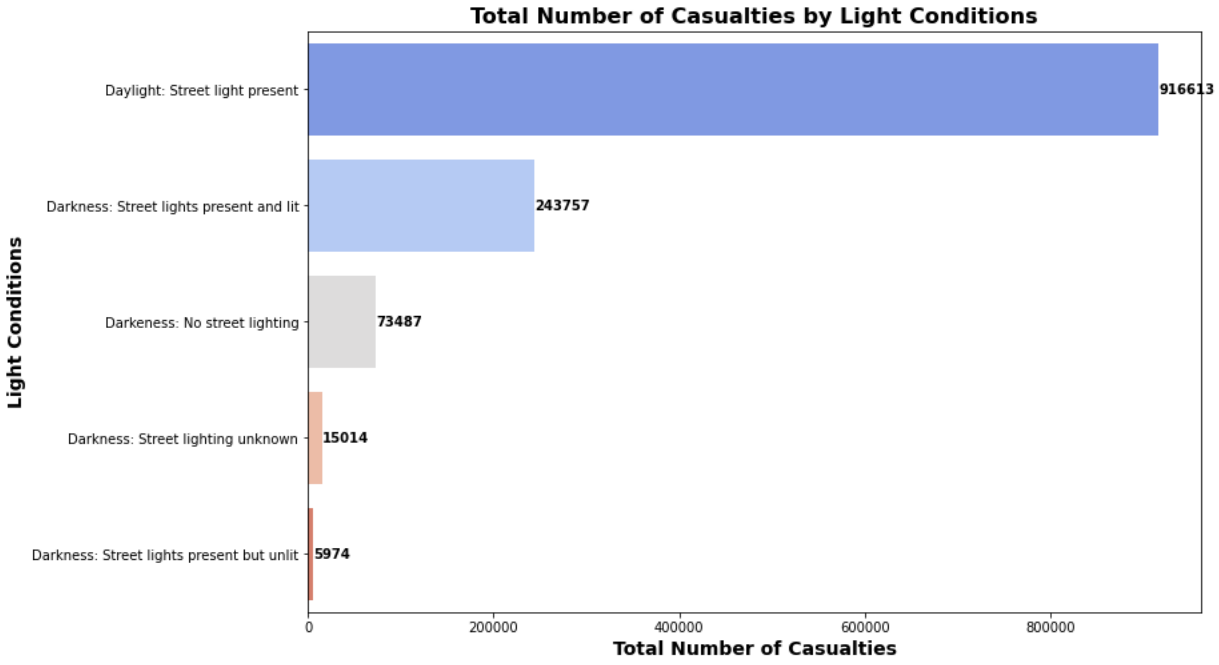

Figure 11

From figure 11, after combining the two datasets with Spark SQL, we are finding the relationship between the lighting conditions and the number of casualties. We first group the lighting condition and find the sum of the number of casualties. Then convert the PySpark data frame to Pandas Date Frame for visualization. Then we plot the graph using bar plot. By this plot we could see that the total number of casualties are more on Daylight when street light is present.

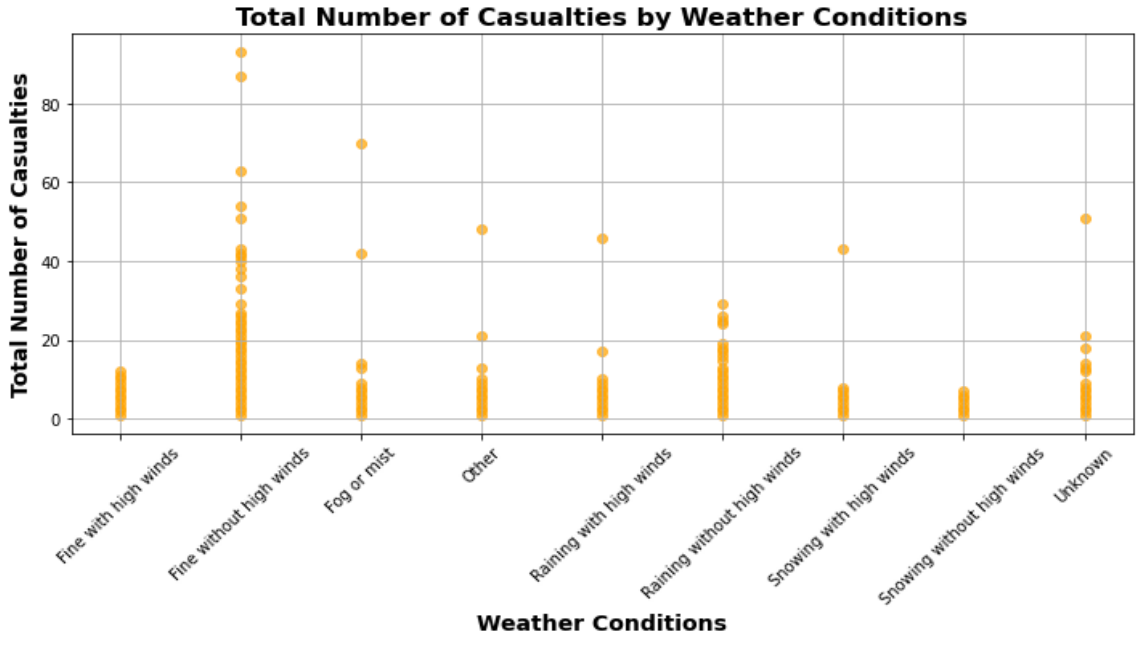

Figure 12

From figure 12, after combining the two datasets with Spark SQL, we are finding the relationship between the weather conditions of our accident dataset and the number of casualties. We first group the weather condition and find the sum of the number of casualties. Then convert the PySpark data frame to Pandas Date Frame for visualization. Then we plot the graph using scatter plot. By this plot we could see that the total number of casualties are more when the weather condition is fine without any high winds and second highest is when fog and mist is present.

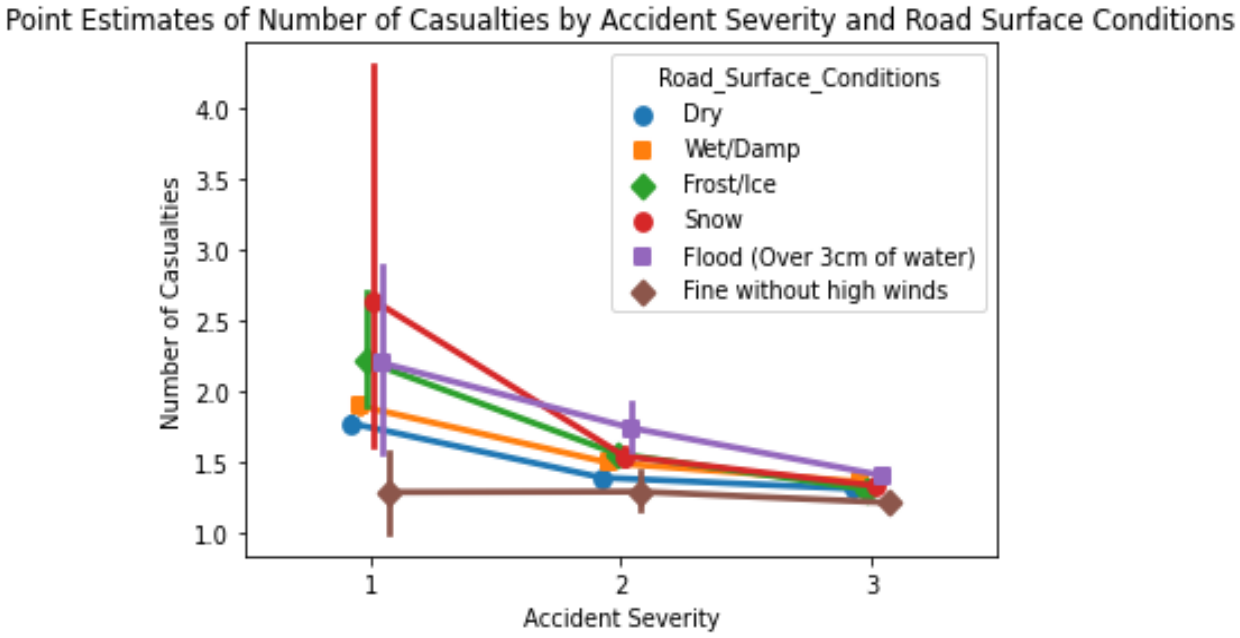

Figure 13

From figure 13, after combining the two datasets with Spark SQL, we are finding the point estimates of number of casualties by accident severity and road surface conditions. We first get the unique categories in road surface conditions using RDD map reduce. Then we plot the graph using point plot. By this plot we could see that the total number of casualties are more on when road surface is Snow where accident severity is 1.

The results largely confirms the risk factors such as speeding, adverse weather, and severe road types. Which indicates that our road safety department should focus their attention to these risk factors when taking action. It is a surprise that in our finding there is a relatively lower impact of lighting conditions on accident rates. This may suggest that drivers are more cautious during low visibility conditions. The correlation between police attendance and data quality was strongly confirmed. This shows that the police plays an important role in road safety and with this involvement more safety measure can be taken to prevent accidents.

F. CONCLUSION AND FUTURE WORK

In this project we conducted a comparative analysis of road accidents from the year 2009 to 2014. By using big data analysis frameworks like Apache Spark and utilizing distributed processing techniques like Map Reduce, we analyzed the large dataset with more than 1 million accident records. We successfully achieved all our research questions such as finding the temporal trends and patterns in road accident over the given time, and we observed that there is a fluctuation in number of accidents happening over the years and it gradually can be decreased with the improved road safety measures. The presence of police officers at accident scenes was found to be an important factor which influence the accuracy and completeness of the accident report without police attendance we cannot get the complete information on how the accident took place. And finally, the key factors like speed limits, weather conditions, road types and light conditions were found that they are the major contributors to the road accidents. Understanding these factors can be useful in improving road safety measures. The data cleaning and preprocessing steps have improved the quality of the data. With the interactive visualizations using libraries like Matplotlib and Seaborn we could clearly understand the data. We could extend our research in future, if we could get additional information on the traffic flow, demographics and vehicle registration numbers, etc. we could also implement the advanced machine learning models to predict the future accident trends and identify the potential high-risk areas.

G. REFERENCES

[1] C. S. Karthikeya Sahith, S. Muppidi and S. Merugula, "Apache Spark Big data Analysis, Performance Tuning, and Spark Application Optimization," *2023 International Conference on Evolutionary Algorithms and Soft Computing Techniques (EASCT)*, Bengaluru, India, 2023, pp. 1-8, doi: 10.1109/EASCT59475.2023.10393086.

[2] M. Junaid, S. A. Wagan, N. M. F. Qureshi, C. S. Nam and D. R. Shin, "Big data Predictive Analytics for Apache Spark using Machine Learning," *2020 Global Conference on Wireless and Optical Technologies (GCWOT)*, Malaga, Spain, 2020, pp. 1-7, doi: 10.1109/GCWOT49901.2020.9391620.

[3] A. G. Shoro and T. R. Soomro, "Big Data Analysis : Apache Spark Perspective Big Data Analysis : Ap Spark Perspective BigDataAnalysisApSparkPerspective," no. JANUARY, 2015.

[4] J M. Armbrust *et al*., "Spark SQL: Relational data processing in spark," *Proc. ACM SIGMOD Int. Conf. Manag. Data*, vol. 2015-May, pp. 1383–1394, 2015, doi:10.1145/2723372.2742797.

[5] M. Assefi, E. Behravesh, G. Liu, and A. P. Tafti, "Big data machine learning using apache spark MLlib," *Proc. – 2017 IEEE Int. Conf. Big Data, Big Data 2017*, vol. 2018-Janua, no. December, pp. 3492–3498, 2017, doi:10.1109/BigData.2017.8258338.