Introduction

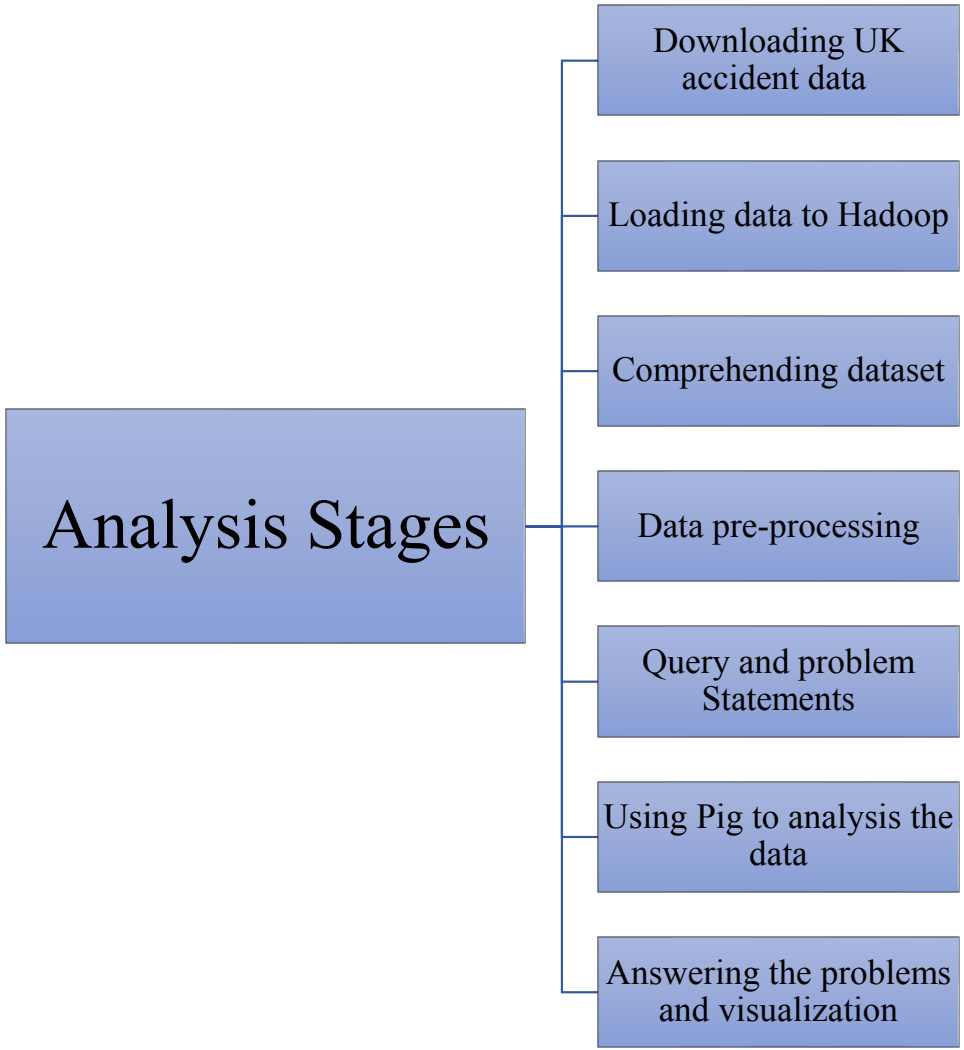
Big Data Analysis of UK Accident by Hadoop in Pig environment

- Based on the statistics data around 1.3 million people are being killed yearly in road accidents and around 40 million people are being injured or experienced disability due to an accident.
- Regarding the considerable economic losses that might happen to individuals, their families, and to nations in road accidents, figuring out the major reasoning of road crashes and finding out the solutions to decrease these losses is a vital task to do.

Two important approaches:

- First: investigating the influence of different conditions such as road type, road surface condition, weather condition and light condition on the severity of accidents.
- Second: finding out the relation between the severity with number of involved vehicles in the accident and impact of the severities on the number of causalities.

In the current project, all the analysis has been done by Pig tool and Python is just used for the visualization.


Analysis Stages
Downloading UK accident data
Loading data to Hadoop
Comprehending dataset
Data pre-processing
Query and problem Statements
Using Pig to analysis the data
Answering the problems and visualization

Architecture

- Software Requirements:
  - Windows OS
  - Virtual Machine to run Hortonworks Sandbox
  - Hadoop
  - Pig
  - Python to do the data visualization
- Hardware Requirements:
  - Hard Disk- at least 500 GB
  - RAM- at least 12 GB
  - Processor- Core i5 or above

Stage 1: Data Source : www.Kaggle.com

Stage 2: Load data to Hadoop

1. Run Hortonworks Sandbox in VM.

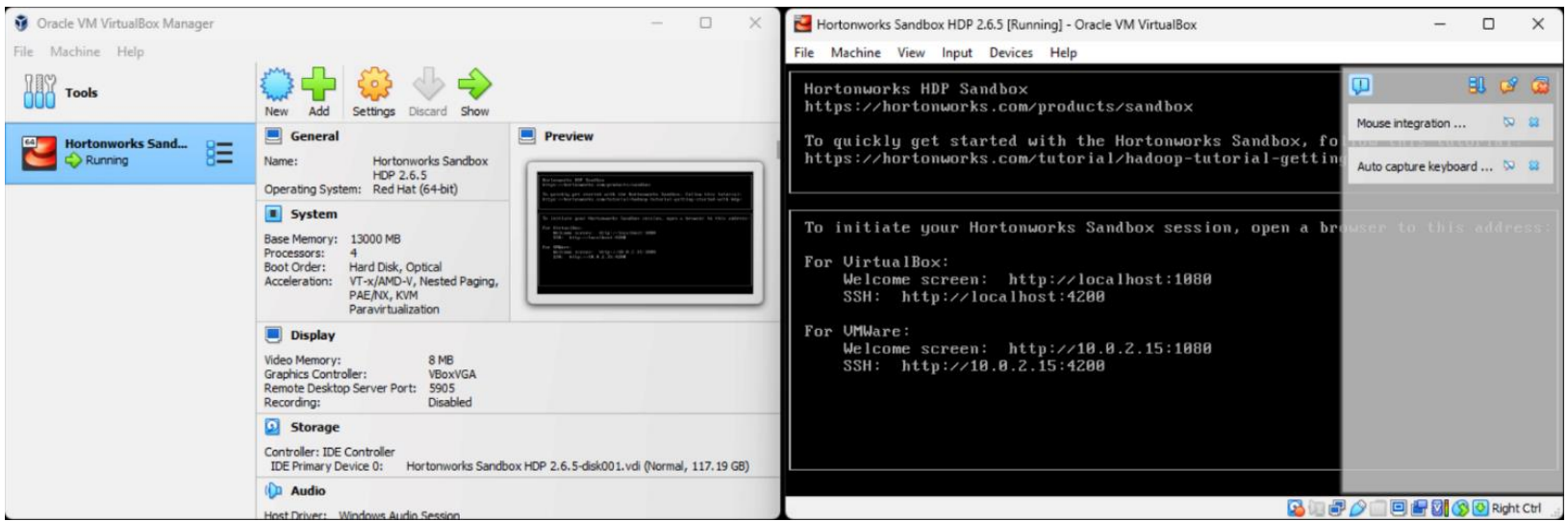


Figure 1

2. Upload data to HDFS.

- Ambari environment → Files View: make a directory to upload the CSV files to HDFS.
- Directory: Big_Data_Project.
- UK_CArAccident_Data.csv is uploaded to HDFS.

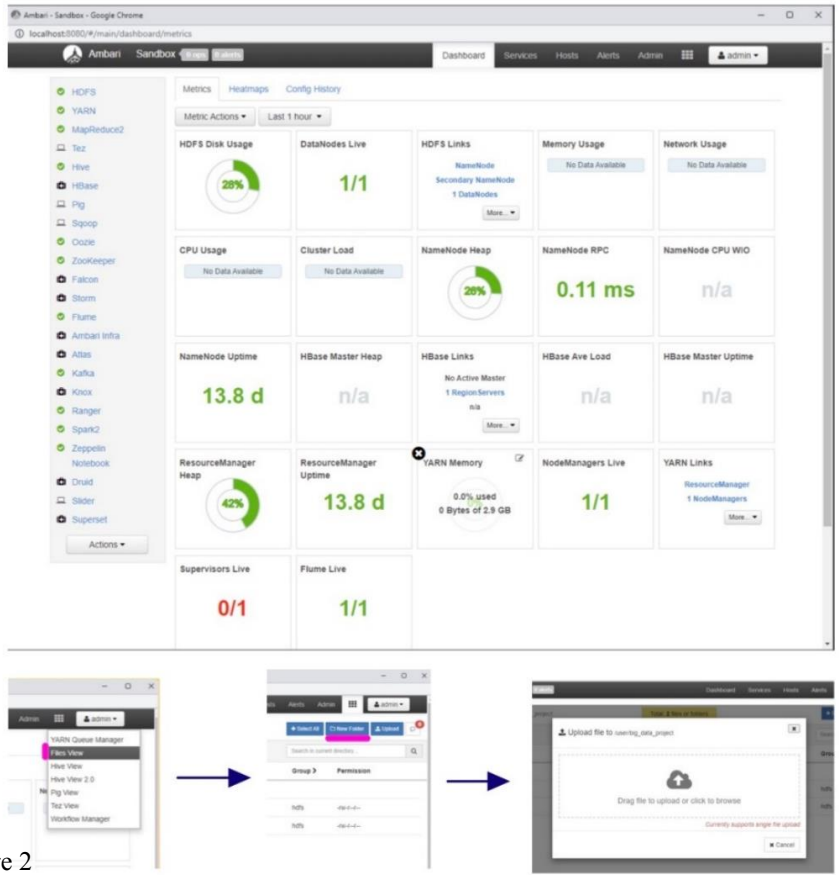


Figure 2

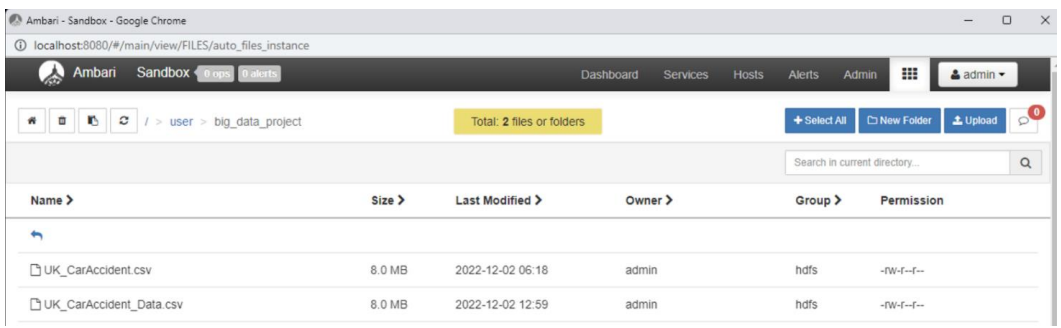


Figure 3

3.Using Pig to analysis the data

- First in the Views Menu ⟶ Pig View
- Pig Interface and using Pig Script to write, run and store the codes and analysis the data.
- 23 script are made for doing different tasks in this project

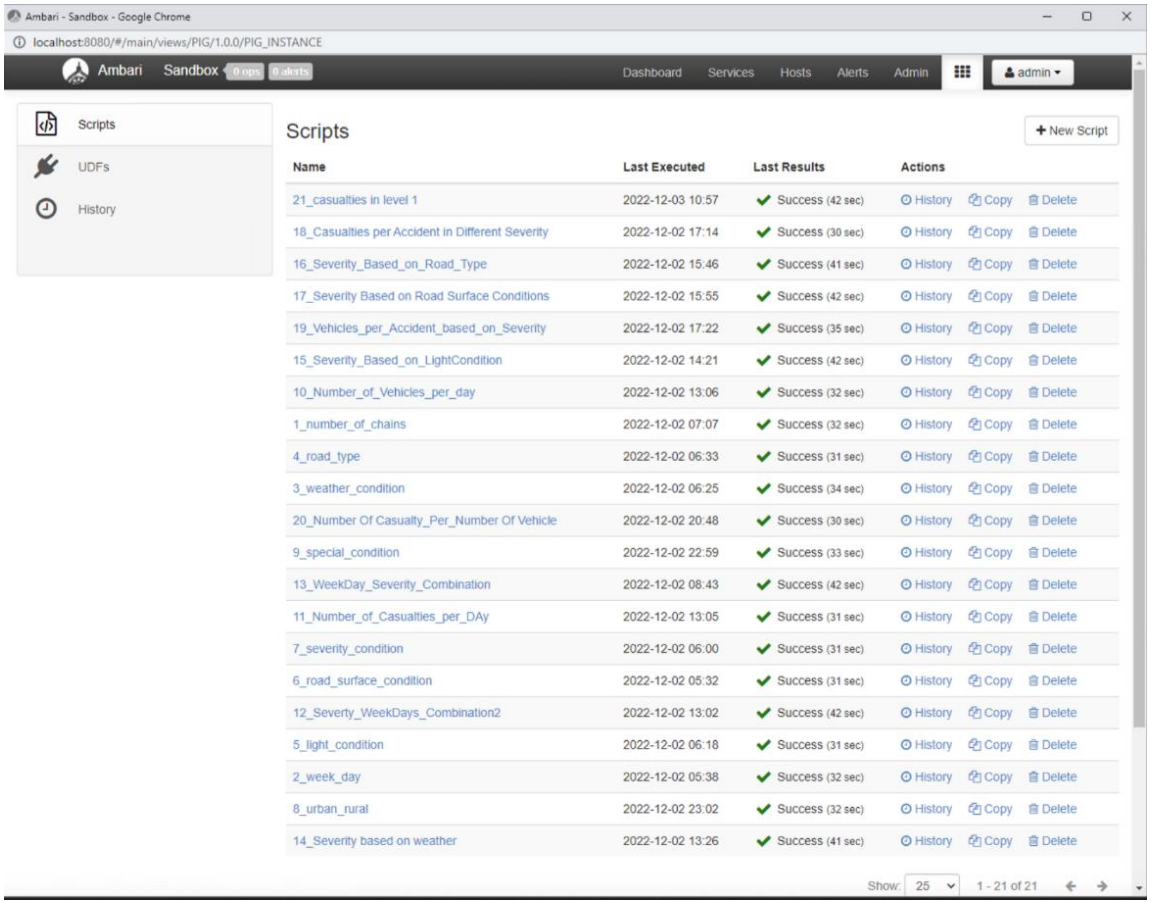


Figure 4

4. Save the result in HDFS

- By STORE command, save the result in HDFS under the "admin/output" directory.
- 31 saved files in the output directory

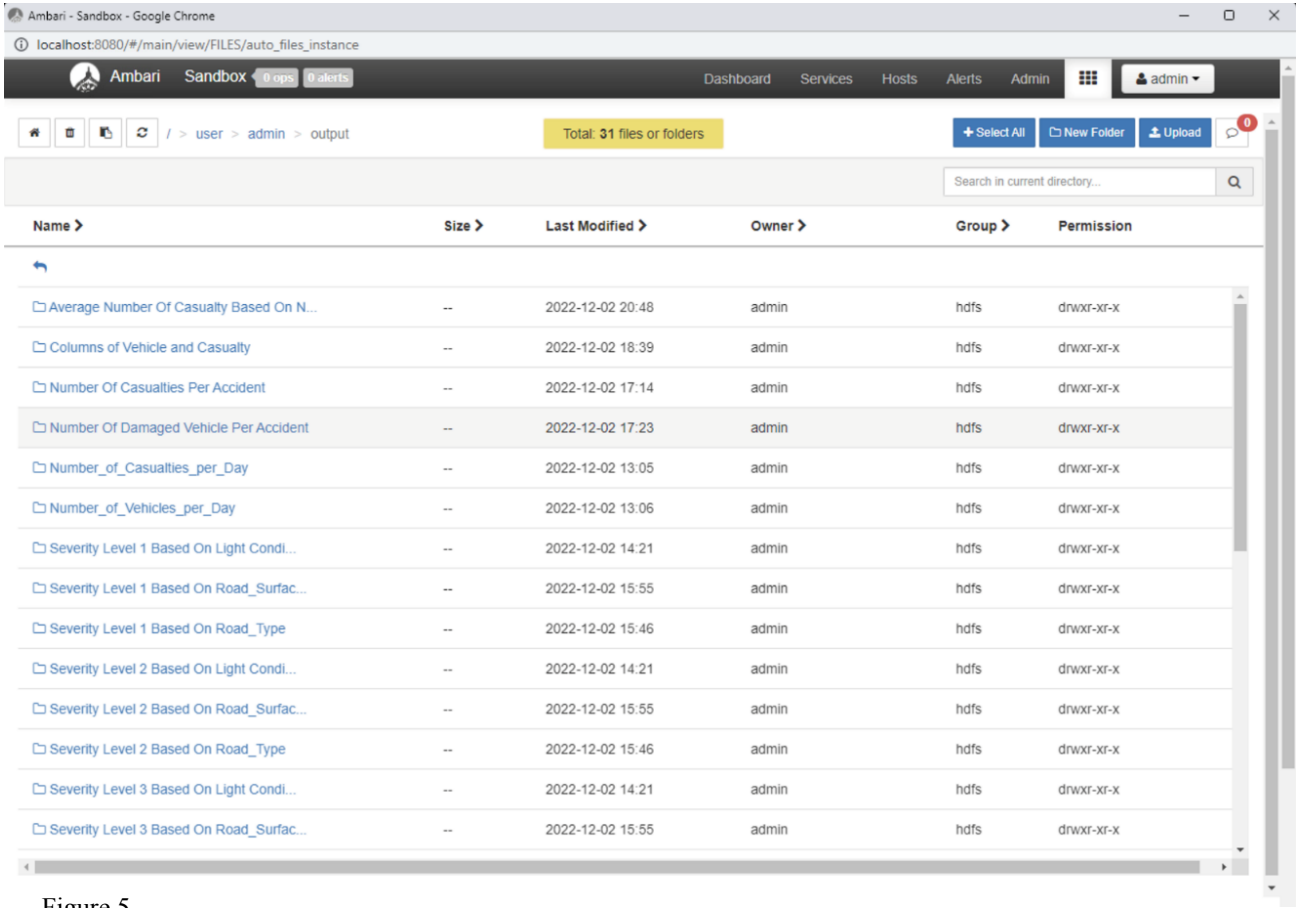


Figure 5

Stage 3: Comprehending datasets

- The UK traffic data from 2000 and 2018
- Over 1.8 million accidents
- 30,000 rows are being considered and analyzed for the current project.
- This dataset consists of 32 various features.

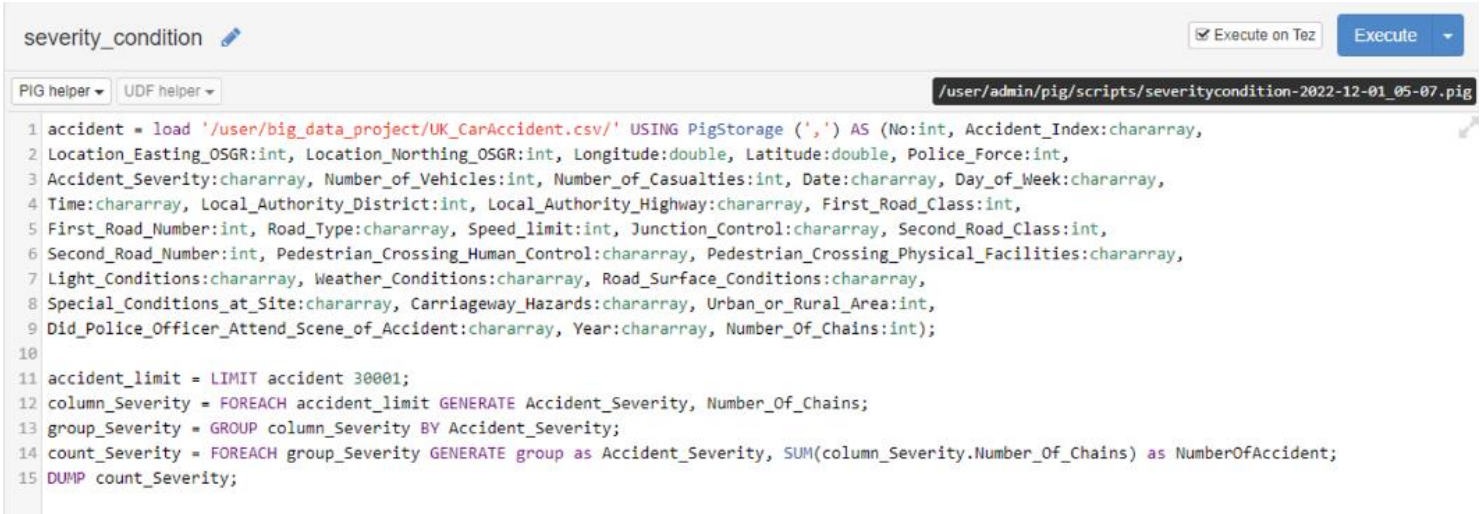
Figure 6

```
df= pd.read_csv('UK_CarAccident.csv')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30032 entries, 0 to 30031
Data columns (total 33 columns):
 #   Column                                       Non-Null Count  Dtype
---  ------                                       --------------  -----
 0   No                                           30032 non-null  int64
 1   Accident_Index                               30032 non-null  object
 2   Location_Easting_OSGR                        30027 non-null  float64
 3   Location_Northing_OSGR                       30032 non-null  int64
 4   Longitude                                    30027 non-null  float64
 5   Latitude                                     30032 non-null  float64
 6   Police_Force                                 30032 non-null  int64
 7   Accident_Severity                            30032 non-null  object
 8   Number_of_Vehicles                           30032 non-null  int64
 9   Number_of_Casualties                         30032 non-null  int64
 10  Date                                         30032 non-null  object
 11  Day_of_Week                                  30032 non-null  int64
 12  Time                                         30031 non-null  object
 13  Local_Authority_(District)                   30032 non-null  int64
 14  Local_Authority_(Highway)                    30032 non-null  object
 15  1st_Road_Class                               30032 non-null  int64
 16  1st_Road_Number                              30032 non-null  int64
 17  Road_Type                                    30032 non-null  object
 18  Speed_limit                                  30032 non-null  int64
 19  Junction_Control                             30032 non-null  object
 20  2nd_Road_Class                               30032 non-null  int64
 21  2nd_Road_Number                              30032 non-null  int64
 22  Pedestrian_Crossing-Human_Control            30029 non-null  object
 23  Pedestrian_Crossing-Physical_Facilities      30030 non-null  object
 24  Daylight conditions                          30032 non-null  object
 25  Weather_Conditions                           30032 non-null  object
 26  Road_Surface_Conditions                      30032 non-null  object
 27  Special_Conditions_at_Site                   30032 non-null  object
 28  Carriageway_Hazards                          30032 non-null  object
 29  Urban_or_Rural_Area                          30032 non-null  int64
 30  Did_Police_Officer_Attend_Scene_of_Accident  30032 non-null  object
 31  Year                                         30032 non-null  int64
 32  Number of chains                             30032 non-null  int64
dtypes: float64(3), int64(15), object(15)
memory usage: 7.6+ MB
```

Description of used features:

1. Accident_Severity (chararray)



1. Accident_Severity (chararray)

- The feature is classified into 5 levels of severity from 1 to 5.
- In our dataset, no accident happened with severity level more than 3.
- The pie chart illustrates that around 87 % of accidents happened in moderate severity (Level 3).

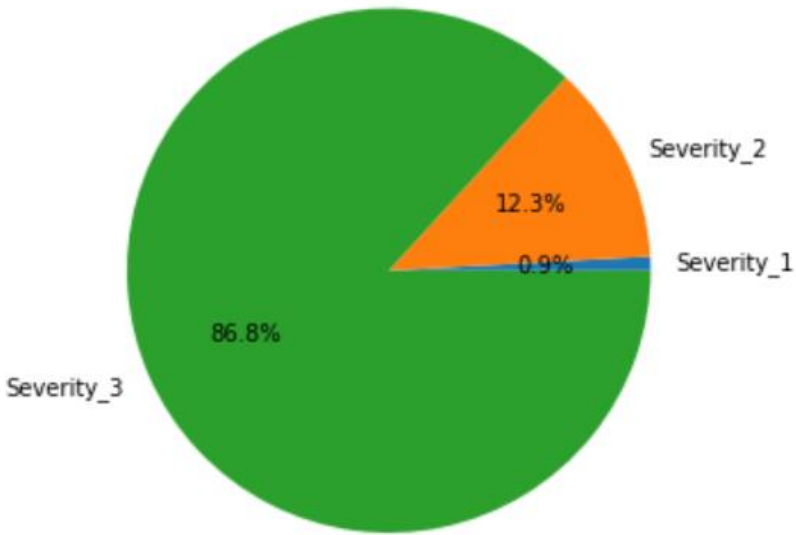
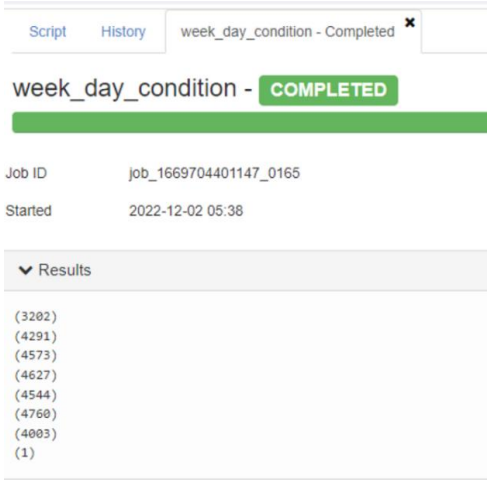
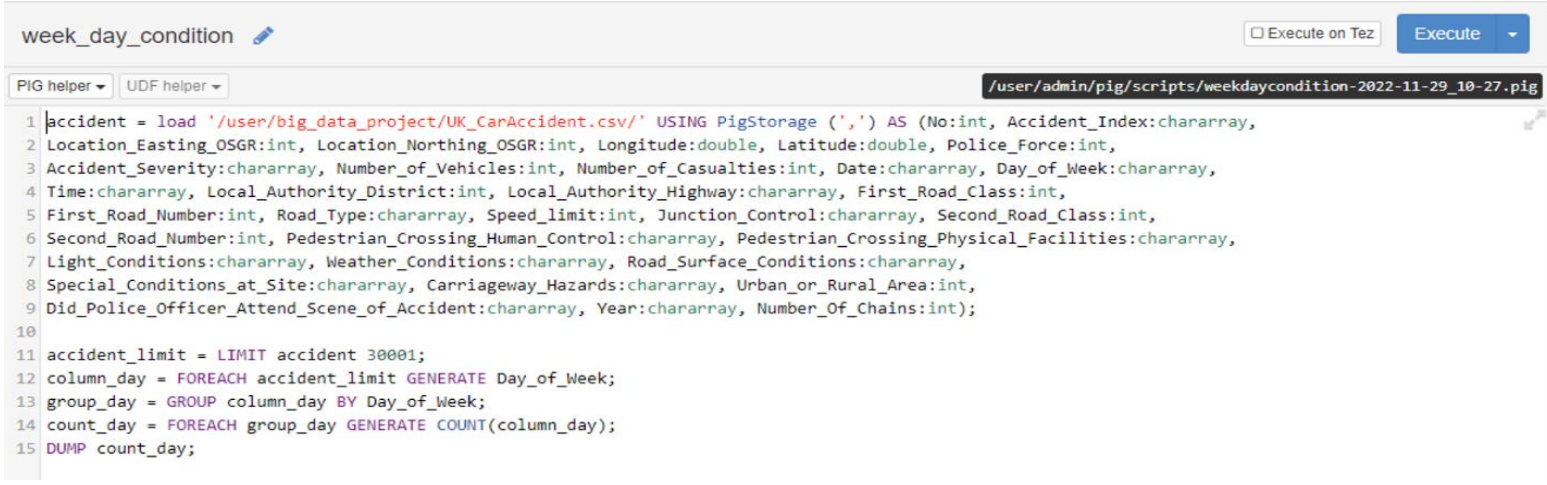


Figure 7

2. Day_of_Week (chararray)





2. Day_of_Week (chararray)

- The feature explains the accidents have happened in which days of the week.
- Number 1,2,3,4,5,6,7 are representing Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday respectively.
- The bar chart reveals the fact that minimum and maximum of accidents have happened on Monday and Saturday respectively.
- In other days, the distribution is almost the same.

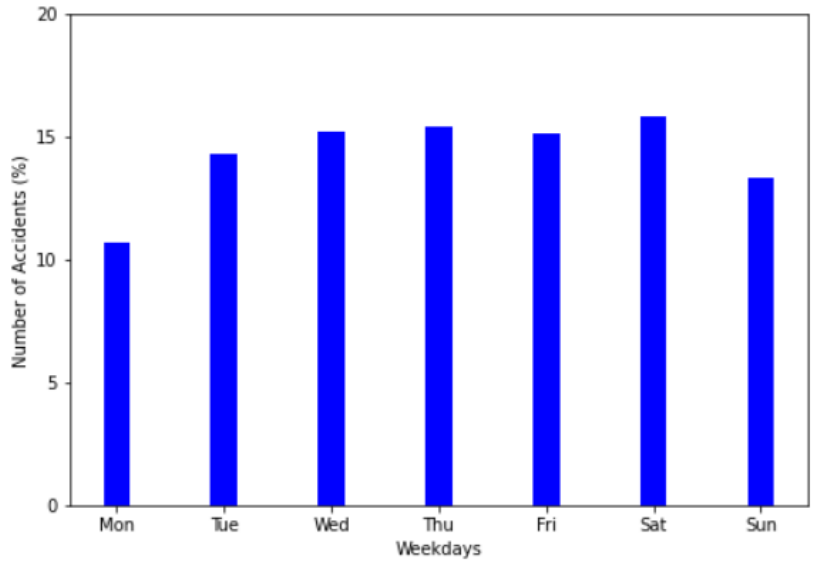


Figure 8

3. Road_Type (Chararray)

road_type - COMPLETED

Job ID	job_1669704401147_0175
Started	2022-12-02 06:33

Results

```
(Unknown,56.0)
(Slip road,162.0)
(Roundabout,1298.0)
(One way street,976.0)
(Dual carriageway,3631.0)
(Single carriageway,23877.0)
```

Script | History

road_type

Execute on Tez | Execute

PIG helper | UDF helper

/user/admin/pig/scripts/roadtype-2022-12-01_10-29.pig

```
1 accident = load '/user/big_data_project/UK_CarAccident.csv/' USING PigStorage (',');
2 accident_N = FILTER accident BY $0>1;
3 accident_limit = LIMIT accident_N 30000;
4 column_road_type = FOREACH accident_limit GENERATE $17 as Road_Type, $32 as Number_Of_Chains;
5 group_road_type = GROUP column_road_type BY Road_Type;
6 count_road_type = FOREACH group_road_type GENERATE group as Road_Type, SUM(column_road_type.Number_Of_Chains) as NumberOfAccident;
7 DUMP count_road_type;
```

3. Road_Type (Chararray)

- This feature is categorized in various road types including Slip Road, Roundabout, One way street, Dual carriageway, Single carriageway.

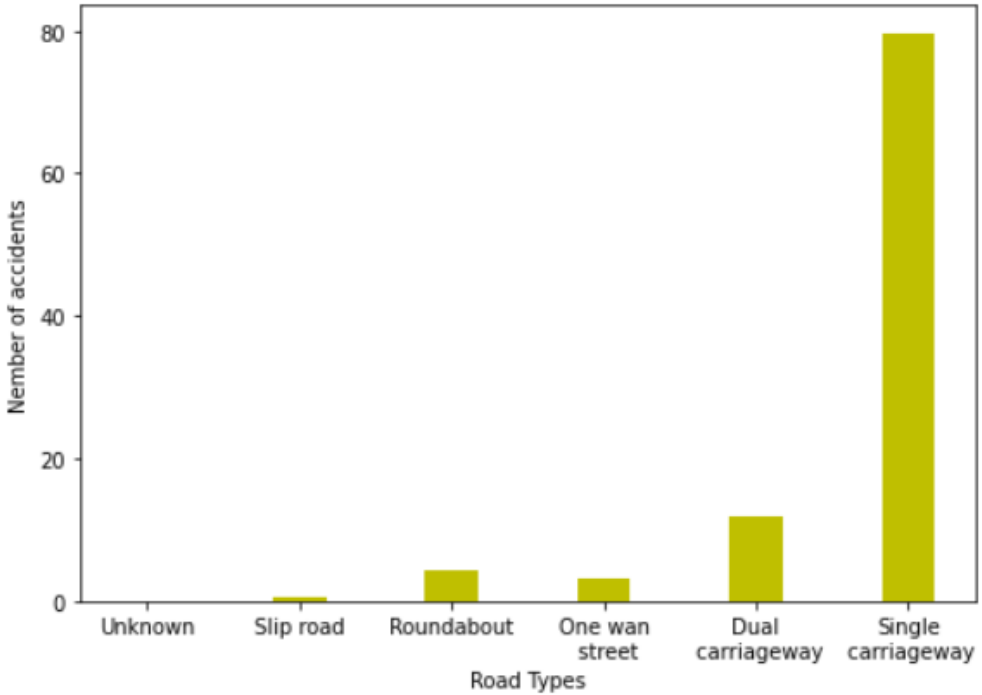


Figure 9

4. Light_Conditions (chararray)

light_condition - COMPLETED

Job ID job_1669704401147_0171

Started 2022-12-02 06:18

Results

```
(Daylight,21179.0)
(Darkness with no street lighting,332.0)
(Darkness with street lighting and lit,8402.0)
(Darkness with street lighting but unlit,87.0)
```

5_light_condition

Execute on Tez | Execute

PIG helper | UDF helper

/user/admin/pig/scripts/lightcondition-2022-11-29_10-27.pig

```
1 accident = load '/user/big_data_project/UK_CarAccident.csv/' USING PigStorage (',');
2 accident_N = FILTER accident BY $0>1;
3 accident_limit = LIMIT accident_N 30000;
4 column_light = FOREACH accident_limit GENERATE $24 as Light_Conditions, $32 as Number_Of_Chains;
5 group_light = GROUP column_light BY Light_Conditions;
6 count_light_condition = FOREACH group_light GENERATE group as Light_Conditions, SUM(column_light.Number_Of_Chains) as NumberOfAccident;
7 DUMP count_light_condition;
8
```

4. Light_Conditions (chararray)

- The feature is described by four categories: Daylight, Darkness with no street lighting, Darkness with street lighting and lit, Darkness with street lighting but unlit.
- The pie chart reveals the fact that two third of accidents were happened during the day and one third during the night.

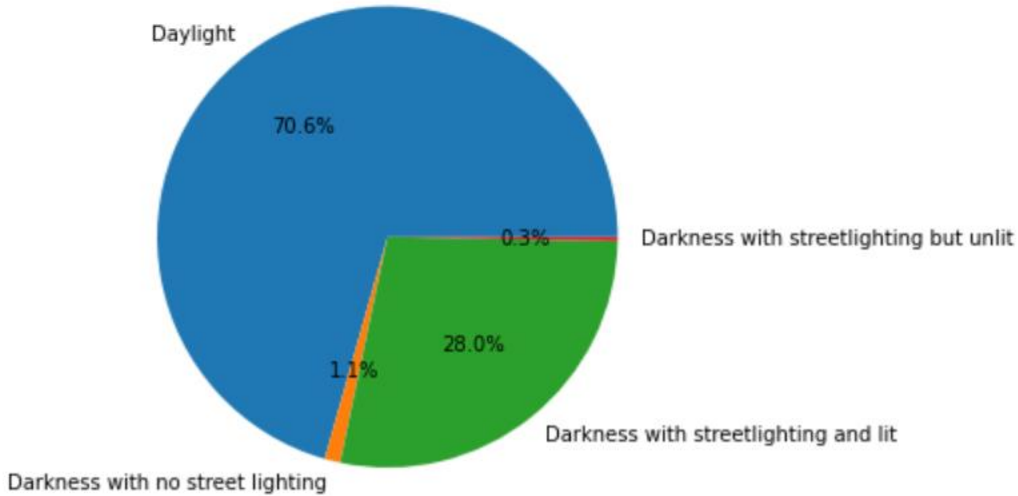


Figure 10

5. Weather_Conditions (chararray)

weather_condition - COMPLETED

Job ID job_1669704401147_0173

Started 2022-12-02 06:25

Results

```
(Other,272.0)
(Unknown,128.0)
(Fog or mist,80.0)
(Fine with high winds,124.0)
(Fine without high winds,25873.0)
(Raining with high winds,91.0)
(Snowing with high winds,11.0)
(Raining without high winds,3267.0)
(Snowing without high winds,154.0)
```

weather_condition

Execute on Tez | Execute

PIG helper | UDF helper

/user/admin/pig/scripts/weathercondition-2022-12-01_10-21.pig

```
1 accident = load '/user/big_data_project/UK_CarAccident.csv/' USING PigStorage (',');
2 accident_N = FILTER accident BY $0>1;
3 accident_limit = LIMIT accident_N 30000;
4 column_weather = FOREACH accident_limit GENERATE $25 as Weather_Conditions, $32 as Number_Of_Chains;
5 group_weather = GROUP column_weather BY Weather_Conditions;
6 count_weather_condition = FOREACH group_weather GENERATE group as Weather_Conditions, SUM(column_weather.Number_Of_Chains) as NumberOfAccident;
7 DUMP count_weather_condition;
```

5. Weather_Conditions (chararray)

- This feature is described by: Fog or mist, Fine with high wind, Fine without high winds, Rain with high winds, Rain without high winds, Snow with high winds and Snow without high winds.
- The bar chart reveals the fact that more than 80 % of accidents happened at a normal weather condition.

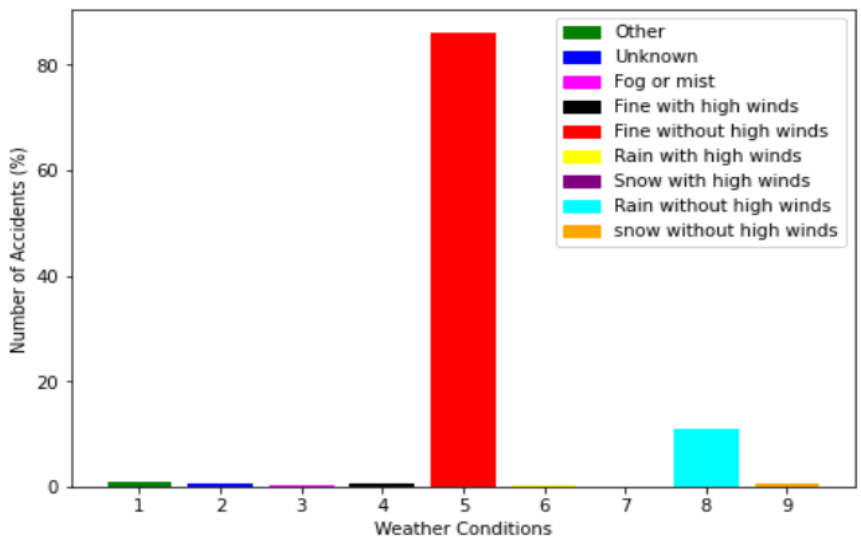
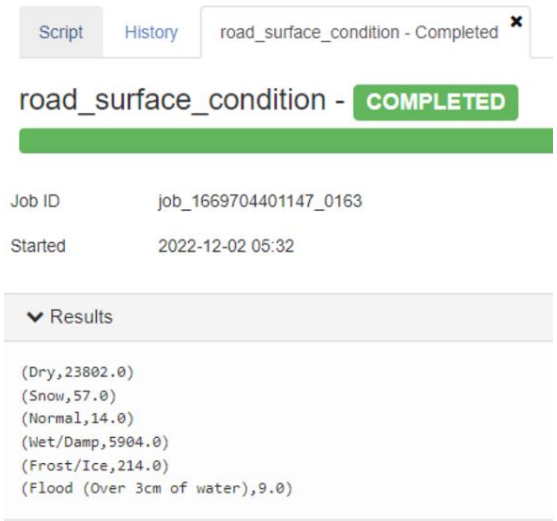


Figure 11

6. Road_Surface_Conditions (chararray)



6. Road_Surface_Conditions (chararray)

- This feature is classified as Dry, Snow, Wet/Damp, Frost/Ice, Flood (Over 3 cm of water).
- The bar chart reveals the fact that 80 % of accidents happened at Dry road surface.

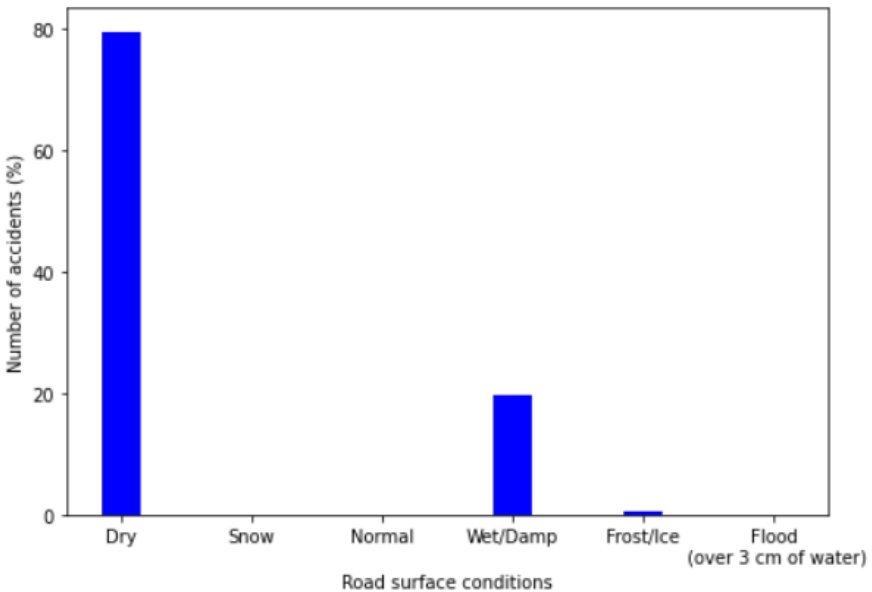
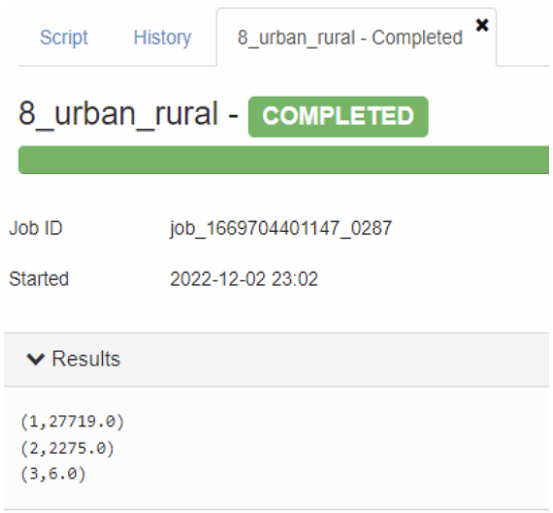


Figure 12

7. Urban_or_Rural_Area (chararray)



7. Urban_or_Rural_Area (chararray)

- This feature is defining the area of the accident. Urban area is represented with number ‘1’ and Rural area by number ‘2’.
- The pie chart shows that the vast majority of accidents had occurred at urban area.

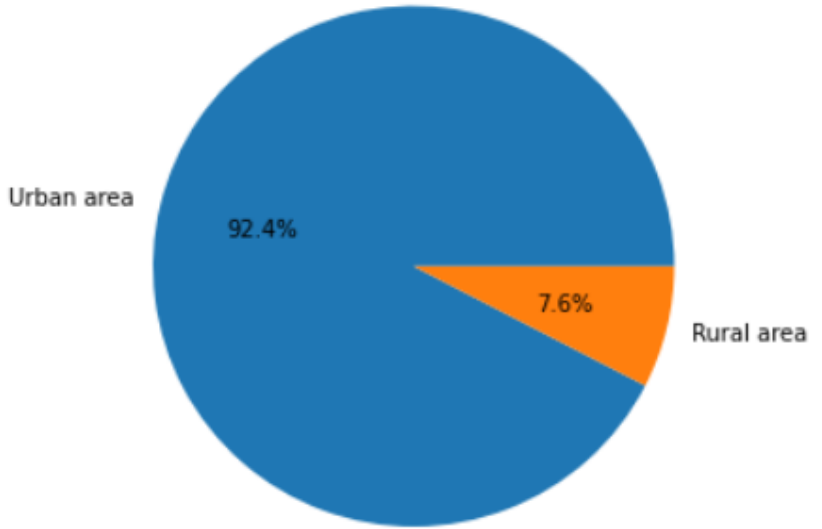
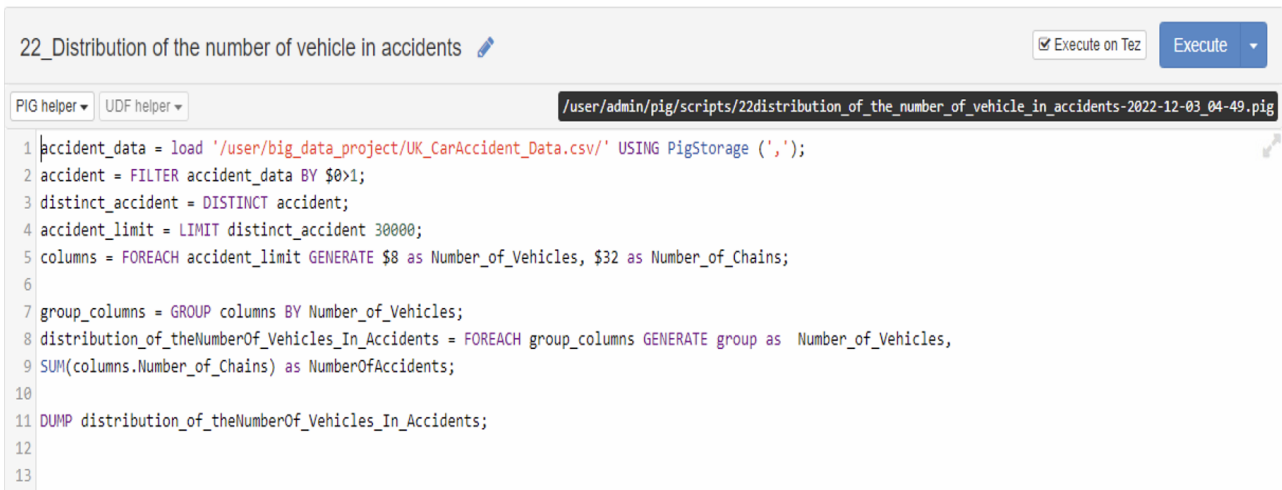


Figure 13

8. Number_of_Vehicles (int)



8. Number_of_Vehicles (int)

- This feature is describing the number of vehicles which are involved in an accident.
- Number of damaged vehicles are varied from 1 to 8.
- In around 60% of accidents, two cars were involved and interestingly in 25% of occasions, the accident happened just by one car.

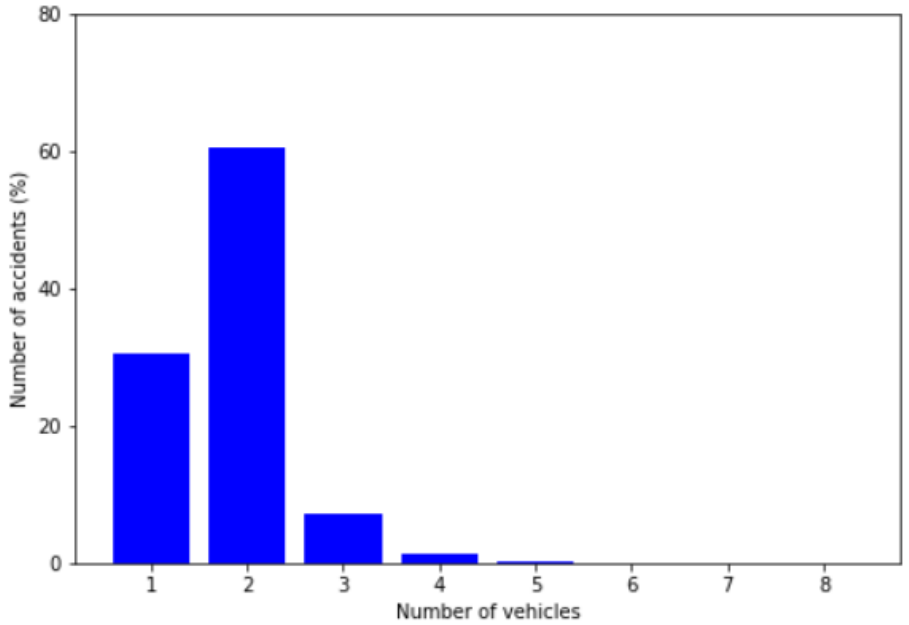


Figure 14

9. Number_of_Causalities (int)

9. Number_of_Causalities (int)

- This feature is describing the number of casualties in an accident.
- The number of casualties is varied from 1 to 23 in different accidents.
- More than 80% of accidents have one person who is injured or killed.

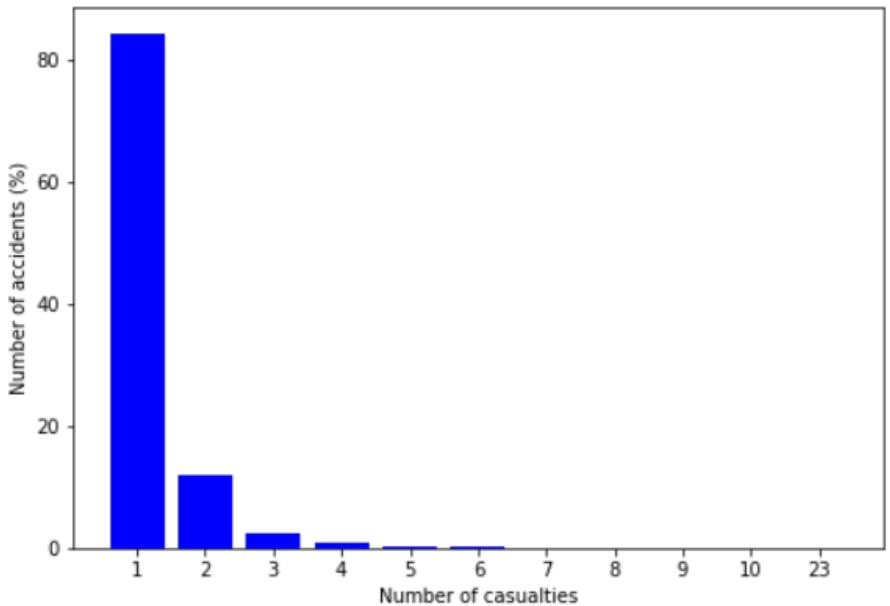
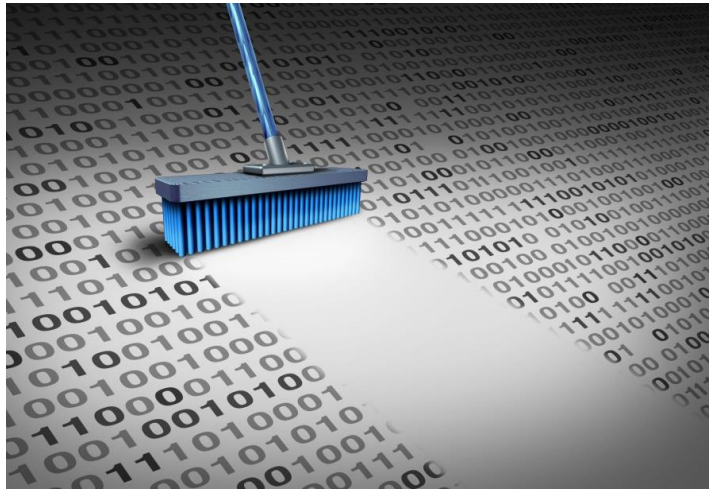


Figure 15

Stage 4: Pre-processing



- In the Road_Type feature, there are 56 accidents happened in the type of the road which is named as ‘unknown’. These data had been added to Single carriageway road type, as 23778 accidents happened in this road type and adding 65 to this number, will not affect the result of our analysis.

- In the Weather_Condition feature, there are two categories defined as “unknown” and “other” with the number of 128 and 272 respectively. They had been added to weather conditions as ‘Fine without high winds” as 25873 accidents in this weather condition would not make change in our analysis.

- In the Road_Surface_Condition feature, one of the conditions is named as “Normal” with 14 accidents. I have added them to the Dry condition with 23802 accidents due to the same reason ad mentioned above.

Stage 4: Query and problem Statements

1. Are the number of accidents, damaged vehicles, and casualties more in weekends than weekdays?
2. In which days of the week, the number of non-serious accidents is less than serious accidents?
3. What is the impact of light conditions on severity of accidents? Are there more accidents happening during nights than days?
4. What is the impact of weather conditions on severity of accidents?
5. Based on Figure 20 most of the accidents had happened Fine weather. Does it mean drivers in rainy or snowy conditions are more cautious?
6. What is the impact of road surface conditions on severity of accidents?
7. Does the wet road surface cause more accidents than dry surface?
8. What is the relation between the road type and the severity of accidents?
9. What is the relation between the severity and the number of cars involved in an accident?
10. How different level of severity affect the number of injured people?

Stages 5 and 6: Using Pig to analysis the data, answering the problems and visualize the result

- To answer the questions and accomplish this section we act as below:
  - Loading the data in Pig environment by LOAD command.
  - Removing the possible replication by DISTINCT command.
  - Decreasing the number of rows to 30,000 by LIMIT command.
  - Continuing the script in a proper way based on each question or problem type and store the result in HDFS by STORE command.
  - Doing the visualization by downloading and using files from HDFS in Python

Q1. Are the number of accidents, damaged vehicles and casualties more in weekends than weekdays?
Q2. In which days of the week, the number of non-serious accidents is less than serious accidents?

Analysis_Severty_WeekDays_Combination2

Execute on Tez | Execute

PIG helper | UDF helper

/user/admin/pig/scripts/severtyweekdayscombination-2022-12-02_12-48.pig

```
accident_data = load '/user/big_data_project/UK_CarAccident_Data.csv/' USING PigStorage (',');
accident = FILTER accident_data BY $0>1;
distinct_accident = DISTINCT accident;
accident_limit = LIMIT distinct_accident 30000;
columns = FOREACH accident_limit GENERATE $7 as Accident_Severity, $11 as Day_of_Week, $32 as Number_Of_Chains;

L1_columns= Filter columns BY Accident_Severity =='Severity_1' ;
L1_final_columns = FOREACH L1_columns GENERATE Day_of_Week, Number_Of_Chains;
L1_group = GROUP L1_final_columns BY Day_of_Week;
L1_count = FOREACH L1_group GENERATE group as Day_of_Week, SUM(L1_final_columns.Number_Of_Chains) as NumberOfAccident;

L2_columns= Filter columns BY Accident_Severity =='Severity_2' ;
L2_final_columns = FOREACH L2_columns GENERATE Day_of_Week, Number_Of_Chains;
L2_group = GROUP L2_final_columns BY Day_of_Week;
L2_count = FOREACH L2_group GENERATE group as Day_of_Week, SUM(L2_final_columns.Number_Of_Chains) as NumberOfAccident;

L3_columns= Filter columns BY Accident_Severity =='Severity_3' ;
L3_final_columns = FOREACH L3_columns GENERATE Day_of_Week, Number_Of_Chains;
L3_group = GROUP L3_final_columns BY Day_of_Week;
L3_count = FOREACH L3_group GENERATE group as Day_of_Week, SUM(L3_final_columns.Number_Of_Chains) as NumberOfAccident;

L4_columns= Filter columns BY Accident_Severity =='Severity_4' ;
L4_final_columns = FOREACH L4_columns GENERATE Day_of_Week, Number_Of_Chains;
L4_group = GROUP L4_final_columns BY Day_of_Week;
L4_count = FOREACH L4_group GENERATE group as Day_of_Week, SUM(L4_final_columns.Number_Of_Chains) as NumberOfAccident;

L5_columns= Filter columns BY Accident_Severity =='Severity_5' ;
L5_final_columns = FOREACH L5_columns GENERATE Day_of_Week, Number_Of_Chains;
L5_group = GROUP L5_final_columns BY Day_of_Week;
L5_count = FOREACH L5_group GENERATE group as Day_of_Week, SUM(L5_final_columns.Number_Of_Chains) as NumberOfAccident;

STORE L1_count INTO 'output/Severity Level1 per Day' using PigStorage(',');
STORE L2_count INTO 'output/Severity Level2 per Day' using PigStorage(',');
STORE L3_count INTO 'output/Severity Level3 per Day' using PigStorage(',');
STORE L4_count INTO 'output/Severity Level4 per Day' using PigStorage(',');
STORE L5_count INTO 'output/Severity Level5 per Day' using PigStorage(',');
```

Script | History | Number_of_Vehicles - Completed

Number_of_Vehicles

Execute on Tez | Execute

PIG helper | UDF helper

/user/admin/pig/scripts/li-2022-12-02_03-20.pig

```
accident_data = load '/user/big_data_project/UK_CarAccident_Data.csv/' USING PigStorage (',');
accident = FILTER accident_data BY $0>1;
distinct_accident = DISTINCT accident;
accident_limit = LIMIT distinct_accident 30000;
columns = FOREACH accident_limit GENERATE $8 as Number_of_Vehicles, $11 as Day_of_Week;

group_columns = GROUP columns BY Day_of_Week;
number_of_vehicles = FOREACH group_columns GENERATE group as Day_of_Week, SUM(columns.Number_of_Vehicles) as NumberOfVehicles;

STORE number_of_vehicles INTO 'output/Number_of_Vehicles_per_Day' using PigStorage(',');
```

File Preview

/user/admin/output/Number_of_Vehicles_per_Day/part-v003-o000-r-00000

```
1,5841.0
2,7871.0
3,8303.0
4,8270.0
5,8132.0
6,8554.0
7,7175.0
```

Script | History | Casualties_per_DAy - Completed

Casualties_per_DAy

Execute on Tez | Execute

PIG helper | UDF helper

/user/admin/pig/scripts/totalseverity-2022-11-30_08-15.pig

```
accident_data = load '/user/big_data_project/UK_CarAccident_Data.csv/' USING PigStorage (',');
accident = FILTER accident_data BY $0>1;
distinct_accident = DISTINCT accident;
accident_limit = LIMIT distinct_accident 30000;
columns = FOREACH accident_limit GENERATE $9 as Number_of_Casualties, $11 as Day_of_Week;

group_columns = GROUP columns BY Day_of_Week;
Number_of_Casualties = FOREACH group_columns GENERATE group as Day_of_Week, SUM(columns.Number_of_Casualties) as NumberOfCasualties;

STORE Number_of_Casualties INTO 'output/Number_of_Casualties_per_Day' using PigStorage(',');
```

File Preview

/user/admin/output/Severity Level1 per Day/part-v005-o000-r-00000

```
1,32.0
2,30.0
3,35.0
4,48.0
5,48.0
6,37.0
7,30.0
```

File Preview

/user/admin/output/Severity Level2 per Day/part-v003-o000-r-00000

```
1,426.0
2,489.0
3,525.0
4,553.0
5,549.0
6,613.0
7,537.0
```

File Preview

/user/admin/output/Severity Level3 per Day/part-v007-o000-r-00000

```
1,2744.0
2,3772.0
3,4012.0
4,4026.0
5,3947.0
6,4111.0
7,3436.0
```

File Preview

/user/admin/output/Number_of_Casualties_per_Day/part-v003-o000-r-00000

```
1,4115.0
2,5135.0
3,5409.0
4,5532.0
5,5413.0
6,5761.0
7,5082.0
```

Q1. Are the number of accidents, damaged vehicles and casualties more in weekends than weekdays?
Q2. In which days of the week, the number of non-serious accidents is less than serious accidents?

• The number of damaged cars and casualties in Saturday and Monday are highest and lowest respectively, but their number is almost the same in weekdays and Sundays.

• Regarding to the severity level in each days, only at Mondays, number of accident with severity level 3 is less than level 1 and 2. So Monday could be considered as a calm day due to more non-severe accidents compare to other days.

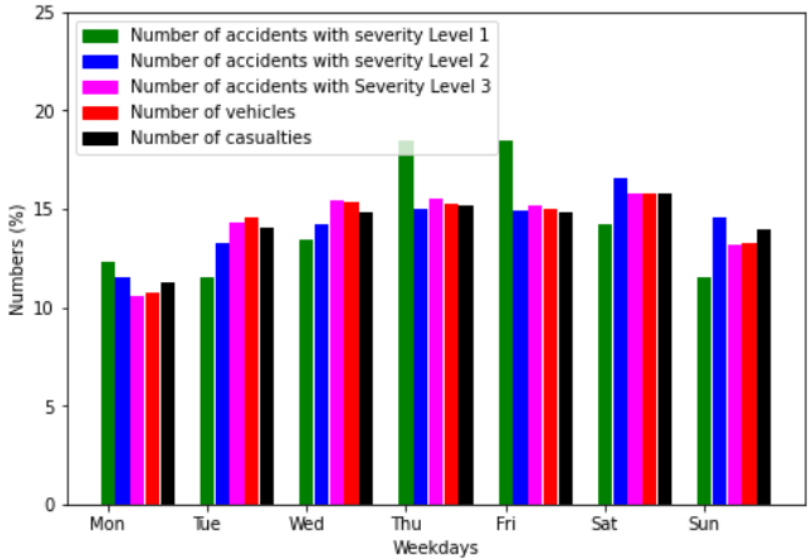


Figure 16. The figure shows the accidents with severity level 1, level 2 and level 3 in different days of the week. Also, the number of vehicles involved in accidents and number of casualties are presented.

Q3. What is the impact of light conditions on severity of accidents? Are there more accidents happening during nights than days?

File Preview

/user/admin/output/Severity Level 1 Based On Light Condition/part-v005-o000-r-00000

```
Daylight,156.0
Darkness with no street lighting,13.0
Darkness with street lighting and lit,90.0
Darkness with street lighting but unlit,1.0
```

File Preview

/user/admin/output/Severity Level 2 Based On Light Condition/part-v003-o000-r-00000

```
Daylight,2429.0
Darkness with no street lighting,82.0
Darkness with street lighting and lit,1171.0
Darkness with street lighting but unlit,10.0
```

File Preview

/user/admin/output/Severity Level 3 Based On Light Condition/part-v007-o000-r-00000

```
Daylight,18594.0
Darkness with no street lighting,237.0
Darkness with street lighting and lit,7141.0
Darkness with street lighting but unlit,76.0
```

14_Severity_Based_on_LightCondition

Execute on Tez | Execute

PIG helper | UDF helper

/user/admin/pig/scripts/14severitybasedonlightcondition-2022-12-02_06-13.pig

```
1  accident_data = load '/user/big_data_project/UK_CarAccident_Data.csv/' USING PigStorage (',');
2  accident = FILTER accident_data BY $0>1;
3  distinct_accident = DISTINCT accident;
4  accident_limit = LIMIT distinct_accident 30000;
5  columns = FOREACH accident_limit GENERATE $7 as Accident_Severity, $24 as Light_Conditions, $32 as Number_Of_Chains;
6
7  L1_columns= Filter columns BY Accident_Severity =='Severity_1' ;
8  L1_final_columns = FOREACH L1_columns GENERATE  Light_Conditions, Number_Of_Chains;
9  L1_group = GROUP L1_final_columns BY  Light_Conditions;
10 L1_count = FOREACH L1_group GENERATE group as Light_Conditions, SUM(L1_final_columns.Number_Of_Chains) as NumberOfAccident;
11
12 L2_columns= Filter columns BY Accident_Severity =='Severity_2' ;
13 L2_final_columns = FOREACH L2_columns GENERATE  Light_Conditions, Number_Of_Chains;
14 L2_group = GROUP L2_final_columns BY  Light_Conditions;
15 L2_count = FOREACH L2_group GENERATE group as Light_Conditions, SUM(L2_final_columns.Number_Of_Chains) as NumberOfAccident;
16
17 L3_columns= Filter columns BY Accident_Severity =='Severity_3' ;
18 L3_final_columns = FOREACH L3_columns GENERATE Light_Conditions, Number_Of_Chains;
19 L3_group = GROUP L3_final_columns BY  Light_Conditions;
20 L3_count = FOREACH L3_group GENERATE group as Light_Conditions, SUM(L3_final_columns.Number_Of_Chains) as NumberOfAccident;
21
22 L4_columns= Filter columns BY Accident_Severity =='Severity_4' ;
23 L4_final_columns = FOREACH L4_columns GENERATE  Light_Conditions, Number_Of_Chains;
24 L4_group = GROUP L4_final_columns BY  Light_Conditions;
25 L4_count = FOREACH L4_group GENERATE group as  Light_Conditions, SUM(L4_final_columns.Number_Of_Chains) as NumberOfAccident;
26
27 L5_columns= Filter columns BY Accident_Severity =='Severity_5' ;
28 L5_final_columns = FOREACH L5_columns GENERATE  Light_Conditions, Number_Of_Chains;
29 L5_group = GROUP L5_final_columns BY  Light_Conditions;
30 L5_count = FOREACH L5_group GENERATE group as  Light_Conditions, SUM(L5_final_columns.Number_Of_Chains) as NumberOfAccident;
31
32 STORE L1_count INTO 'output/Severity Level 1 Based On Light Condition' using PigStorage(',');
33 STORE L2_count INTO 'output/Severity Level 2 Based On Light Condition' using PigStorage(',');
34 STORE L3_count INTO 'output/Severity Level 3 Based On Light Condition' using PigStorage(',');
35 STORE L4_count INTO 'output/Severity Level 4 Based On Light Condition' using PigStorage(',');
36 STORE L5_count INTO 'output/Severity Level 5 Based On Light Condition' using PigStorage(',');
```

Q3. What is the impact of light conditions on severity of accidents? Are there more accidents happening during nights than days?

- Even though, the number of accidents in daylight is strongly higher than the night, the light condition does not affect the level of severity.
- Even the severity level 3 in darkness without any lighting, is less than other situations.

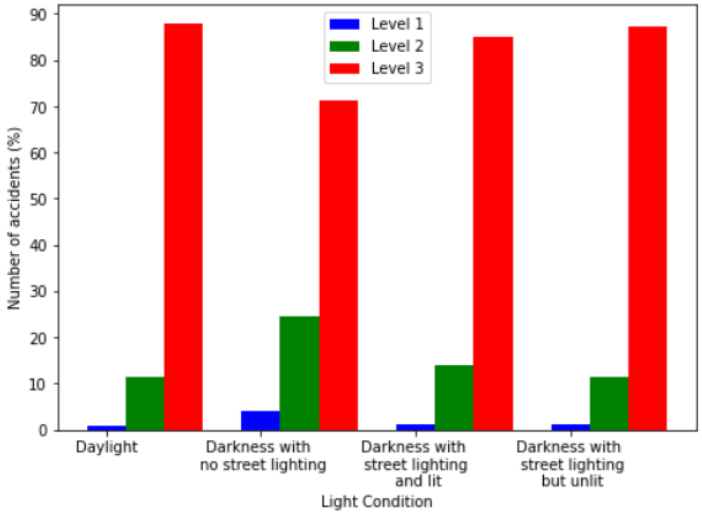


Figure 17. The bar chart represents the number of accidents with different severity levels in different light conditions.

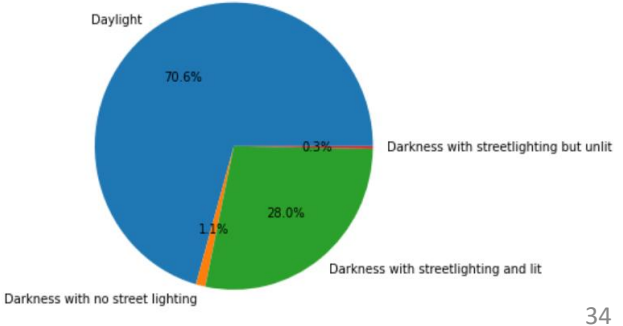


Figure 10

Q4. What is the impact of weather conditions on severity of accidents?

Q5. Based on Figure 20 most of the accidents had happened Fine weather. Does it mean drivers in rainy or snowy conditions are more cautious?

File Preview

/user/admin/output/Severity Level1 Based On Weather/part-v005-o000-r-00000

```
Fog or mist,1.0
Fine with high winds,6.0
Fine without high winds,226.0
Raining with high winds,1.0
Raining without high winds,26.0
```

File Preview

/user/admin/output/Severity Level3 Based On Weather/part-v007-o000-r-00000

```
Fog or mist,65.0
Fine with high winds,108.0
Fine without high winds,22797.0
Raining with high winds,73.0
Snowing with high winds,8.0
Raining without high winds,2857.0
Snowing without high winds,140.0
```

File Preview

/user/admin/output/Severity Level2 Based On Weather/part-v003-o000-r-00000

```
Fog or mist,14.0
Fine with high winds,10.0
Fine without high winds,3250.0
Raining with high winds,17.0
Snowing with high winds,3.0
Raining without high winds,384.0
Snowing without high winds,14.0
```

Severity based on weather

Execute on Tez | Execute

PIG helper | UDF helper

/user/admin/pig/scripts/car-2022-11-23_02-35.pig

```
accident_data = load '/user/big_data_project/UK_CarAccident_Data.csv/' USING PigStorage (',');
accident = FILTER accident_data BY $0>1;
distinct_accident = DISTINCT accident;
accident_limit = LIMIT distinct_accident 30000;
columns = FOREACH accident_limit GENERATE $7 as Accident_Severity, $25 as Weather_Conditions, $32 as Number_Of_Chains;

L1_columns= Filter columns BY Accident_Severity =='Severity_1' ;
L1_final_columns = FOREACH L1_columns GENERATE  Weather_Conditions, Number_Of_Chains;
L1_group = GROUP L1_final_columns BY  Weather_Conditions;
L1_count = FOREACH L1_group GENERATE group as Weather_Conditions, SUM(L1_final_columns.Number_Of_Chains) as NumberOfAccident;

L2_columns= Filter columns BY Accident_Severity =='Severity_2' ;
L2_final_columns = FOREACH L2_columns GENERATE  Weather_Conditions, Number_Of_Chains;
L2_group = GROUP L2_final_columns BY  Weather_Conditions;
L2_count = FOREACH L2_group GENERATE group as Weather_Conditions, SUM(L2_final_columns.Number_Of_Chains) as NumberOfAccident;

L3_columns= Filter columns BY Accident_Severity =='Severity_3' ;
L3_final_columns = FOREACH L3_columns GENERATE Weather_Conditions, Number_Of_Chains;
L3_group = GROUP L3_final_columns BY  Weather_Conditions;
L3_count = FOREACH L3_group GENERATE group as Weather_Conditions, SUM(L3_final_columns.Number_Of_Chains) as NumberOfAccident;

L4_columns= Filter columns BY Accident_Severity =='Severity_4' ;
L4_final_columns = FOREACH L4_columns GENERATE  Weather_Conditions, Number_Of_Chains;
L4_group = GROUP L4_final_columns BY  Weather_Conditions;
L4_count = FOREACH L4_group GENERATE group as  Weather_Conditions, SUM(L4_final_columns.Number_Of_Chains) as NumberOfAccident;

L5_columns= Filter columns BY Accident_Severity =='Severity_5' ;
L5_final_columns = FOREACH L5_columns GENERATE  Weather_Conditions, Number_Of_Chains;
L5_group = GROUP L5_final_columns BY  Weather_Conditions;
L5_count = FOREACH L5_group GENERATE group as  Weather_Conditions, SUM(L5_final_columns.Number_Of_Chains) as NumberOfAccident;

STORE L1_count INTO 'output/Severity Level1 Based On Weather' using PigStorage(',');
STORE L2_count INTO 'output/Severity Level2 Based On Weather' using PigStorage(',');
STORE L3_count INTO 'output/Severity Level3 Based On Weather' using PigStorage(',');
STORE L4_count INTO 'output/Severity Level4 Based On Weather' using PigStorage(',');
STORE L5_count INTO 'output/Severity Level5 Based On Weather' using PigStorage(',');
```

Q4. What is the impact of weather conditions on severity of accidents?
Q5. Based on Figure 20 most of the accidents had happened Fine weather. Does it mean drivers in rainy or snowy conditions are more cautious?

- Although almost 90 per cent of accidents had happened in fine weather, it doesn't mean that drivers in rainy days are more careful.
- The Figure reveals the most sever accident happened in fog or mist, which means the drivers were not cautious enough in that condition.
- Furthermore, there are minor differences between the number of severity level 3 accidents in various situations..

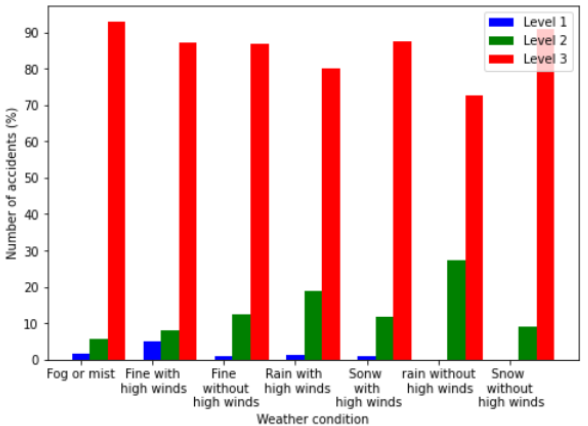


Figure 18. The bar chart represents the number of accidents with different severity levels in different weather conditions.

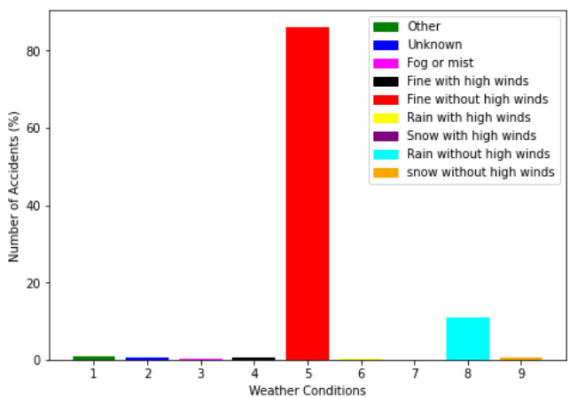


Figure 11

Q6. What is the impact of road surface conditions on severity of accidents?
Q7. Does the wet road surface cause more accidents than dry surface?

File Preview

/user/admin/output/Severity Level 1 Based On Road_Surface_Conditions/part-v005-o000-r-00000

```
Dry,190.0
Wet/Damp,67.0
Frost/Ice,3.0
```

File Preview

/user/admin/output/Severity Level 2 Based On Road_Surface_Conditions/part-v003-o000-r-00000

```
Dry,2918.0
Snow,10.0
Wet/Damp,735.0
Frost/Ice,28.0
Flood (Over 3cm of water),1.0
```

File Preview

/user/admin/output/Severity Level 3 Based On Road_Surface_Conditions/part-v007-o000-r-00000

```
Dry,20708.0
Snow,47.0
Wet/Damp,5102.0
Frost/Ice,183.0
Flood (Over 3cm of water),8.0
```

17_Severity Based on Road Surface Conditions

Execute on Tez | Execute

PIG helper | UDF helper

/user/admin/pig/scripts/17severity_based_on_road_surface_conditions_-2022-12-02_07-54.pig

```
1  accident_data = load '/user/big_data_project/UK_CarAccident_Data.csv/' USING PigStorage (',');
2  accident = FILTER accident_data BY $0>1;
3  distinct_accident = DISTINCT accident;
4  accident_limit = LIMIT distinct_accident 30000;
5  columns = FOREACH accident_limit GENERATE $7 as Accident_Severity, $26 as Road_Surface_Conditions, $32 as Number_Of_Chains;
6
7  L1_columns= Filter columns BY Accident_Severity =='Severity_1' ;
8  L1_final_columns = FOREACH L1_columns GENERATE  Road_Surface_Conditions, Number_Of_Chains;
9  L1_group = GROUP L1_final_columns BY  Road_Surface_Conditions;
10 L1_count = FOREACH L1_group GENERATE group as Road_Surface_Conditions, SUM(L1_final_columns.Number_Of_Chains) as NumberOfAccident;
11
12 L2_columns= Filter columns BY Accident_Severity =='Severity_2' ;
13 L2_final_columns = FOREACH L2_columns GENERATE Road_Surface_Conditions, Number_Of_Chains;
14 L2_group = GROUP L2_final_columns BY Road_Surface_Conditions;
15 L2_count = FOREACH L2_group GENERATE group as Road_Surface_Conditions, SUM(L2_final_columns.Number_Of_Chains) as NumberOfAccident;
16
17 L3_columns= Filter columns BY Accident_Severity =='Severity_3' ;
18 L3_final_columns = FOREACH L3_columns GENERATE Road_Surface_Conditions, Number_Of_Chains;
19 L3_group = GROUP L3_final_columns BY Road_Surface_Conditions;
20 L3_count = FOREACH L3_group GENERATE group as Road_Surface_Conditions, SUM(L3_final_columns.Number_Of_Chains) as NumberOfAccident;
21
22 L4_columns= Filter columns BY Accident_Severity =='Severity_4' ;
23 L4_final_columns = FOREACH L4_columns GENERATE Road_Surface_Conditions, Number_Of_Chains;
24 L4_group = GROUP L4_final_columns BY  Road_Surface_Conditions;
25 L4_count = FOREACH L4_group GENERATE group as  Road_Surface_Conditions, SUM(L4_final_columns.Number_Of_Chains) as NumberOfAccident;
26
27 L5_columns= Filter columns BY Accident_Severity =='Severity_5' ;
28 L5_final_columns = FOREACH L5_columns GENERATE  Road_Surface_Conditions, Number_Of_Chains;
29 L5_group = GROUP L5_final_columns BY Road_Surface_Conditions;
30 L5_count = FOREACH L5_group GENERATE group as Road_Surface_Conditions, SUM(L5_final_columns.Number_Of_Chains) as NumberOfAccident;
31
32 STORE L1_count INTO 'output/Severity Level 1 Based On Road_Surface_Conditions' using PigStorage(',');
33 STORE L2_count INTO 'output/Severity Level 2 Based On Road_Surface_Conditions' using PigStorage(',');
34 STORE L3_count INTO 'output/Severity Level 3 Based On Road_Surface_Conditions' using PigStorage(',');
35 STORE L4_count INTO 'output/Severity Level 4 Based On Road_Surface_Conditions' using PigStorage(',');
36 STORE L5_count INTO 'output/Severity Level 5 Based On Road_Surface_Conditions' using PigStorage(',');
```

Q6. What is the impact of road surface conditions on severity of accidents?

Q7. Does the wet road surface cause more accidents than dry surface?

- The severity level in each road surface condition is almost the same, however in flood situation, severity level 3 is higher than others. It may represent that people do not speed up in not proper road surfaces.
- The number of accidents is higher in dry road than wet road mainly because most of the days of the year are sunny.

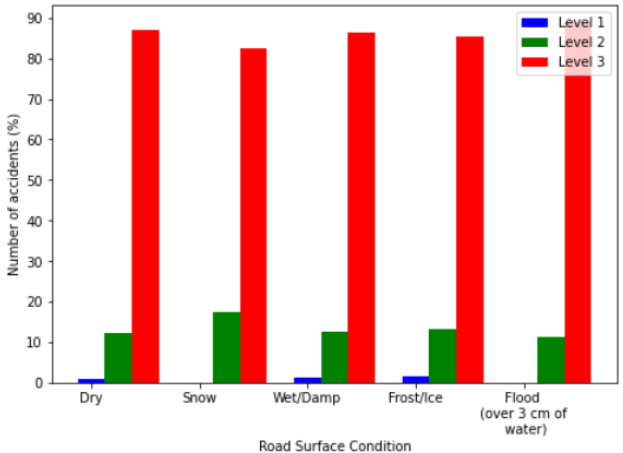


Figure 19. The bar chart represents the number of accidents with different severity levels in different road surface conditions

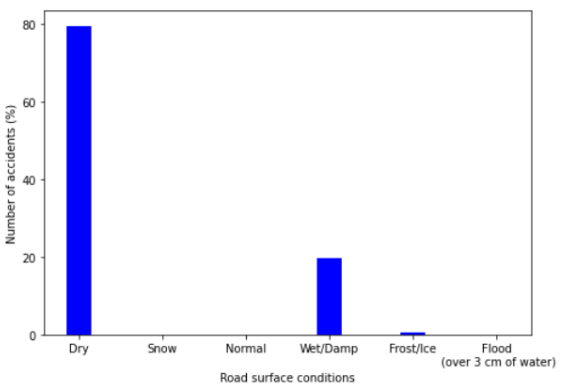


Figure 12

Q8. What is the relation between the road type and the severity of accidents?

File Preview

/user/admin/output/Severity Level 1 Based On Road_Type/part-v005-o000-r-00000

```
Slip road,2.0
Roundabout,6.0
One way street,5.0
Dual carriageway,44.0
Single carriageway,203.0
```

File Preview

/user/admin/output/Severity Level 2 Based On Road_Type/part-v003-o000-r-00000

```
Slip road,17.0
Roundabout,101.0
One way street,125.0
Dual carriageway,497.0
Single carriageway,2952.0
```

File Preview

/user/admin/output/Severity Level 3 Based On Road_Type/part-v007-o000-r-00000

```
Slip road,143.0
Roundabout,1191.0
One way street,846.0
Dual carriageway,3090.0
Single carriageway,20778.0
```

16_Severity_Based_on_Road_Type

Execute on Tez | Execute

PIG helper | UDF helper

/user/admin/pig/scripts/16severitybasedonroadtype-2022-12-02_06-34.pig

```
accident_data = load '/user/big_data_project/UK_CarAccident_Data.csv/' USING PigStorage (',');
accident = FILTER accident_data BY $0>1;
distinct_accident = DISTINCT accident;
accident_limit = LIMIT distinct_accident 30000;
columns = FOREACH accident_limit GENERATE $7 as Accident_Severity, $17 as Road_Type, $32 as Number_Of_Chains;

L1_columns= Filter columns BY Accident_Severity =='Severity_1' ;
L1_final_columns = FOREACH L1_columns GENERATE  Road_Type, Number_Of_Chains;
L1_group = GROUP L1_final_columns BY  Road_Type;
L1_count = FOREACH L1_group GENERATE group as Road_Type, SUM(L1_final_columns.Number_Of_Chains) as NumberOfAccident;

L2_columns= Filter columns BY Accident_Severity =='Severity_2' ;
L2_final_columns = FOREACH L2_columns GENERATE Road_Type, Number_Of_Chains;
L2_group = GROUP L2_final_columns BY Road_Type;
L2_count = FOREACH L2_group GENERATE group as Road_Type, SUM(L2_final_columns.Number_Of_Chains) as NumberOfAccident;

L3_columns= Filter columns BY Accident_Severity =='Severity_3' ;
L3_final_columns = FOREACH L3_columns GENERATE Road_Type, Number_Of_Chains;
L3_group = GROUP L3_final_columns BY Road_Type;
L3_count = FOREACH L3_group GENERATE group as Road_Type, SUM(L3_final_columns.Number_Of_Chains) as NumberOfAccident;

L4_columns= Filter columns BY Accident_Severity =='Severity_4' ;
L4_final_columns = FOREACH L4_columns GENERATE  Road_Type, Number_Of_Chains;
L4_group = GROUP L4_final_columns BY  Road_Type;
L4_count = FOREACH L4_group GENERATE group as  Road_Type, SUM(L4_final_columns.Number_Of_Chains) as NumberOfAccident;

L5_columns= Filter columns BY Accident_Severity =='Severity_5' ;
L5_final_columns = FOREACH L5_columns GENERATE  Road_Type, Number_Of_Chains;
L5_group = GROUP L5_final_columns BY  Road_Type;
L5_count = FOREACH L5_group GENERATE group as Road_Type, SUM(L5_final_columns.Number_Of_Chains) as NumberOfAccident;

STORE L1_count INTO 'output/Severity Level 1 Based On Road_Type' using PigStorage(',');
STORE L2_count INTO 'output/Severity Level 2 Based On Road_Type' using PigStorage(',');
STORE L3_count INTO 'output/Severity Level 3 Based On Road_Type' using PigStorage(',');
STORE L4_count INTO 'output/Severity Level 4 Based On Road_Type' using PigStorage(',');
STORE L5_count INTO 'output/Severity Level 5 Based On Road_Type' using PigStorage(',');
```

Q8. What is the relation between the road type and the severity of accidents?

- More severe accidents happened in roundabouts and slip roads.
- Dual carriageways are caused the less severe accidents compare to single carriageway and one way street.

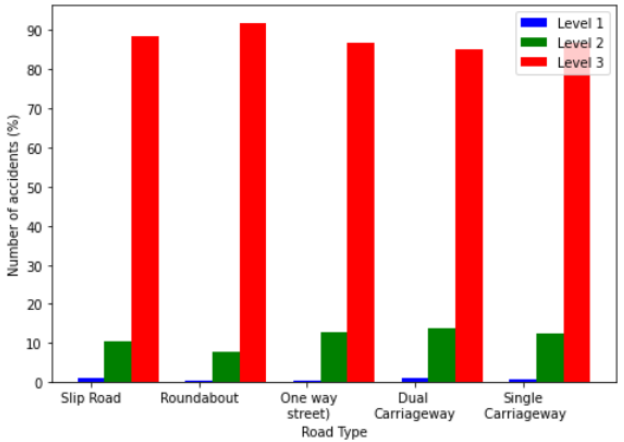


Figure 20. The bar chart represents the number of accidents with different severity levels in different road types

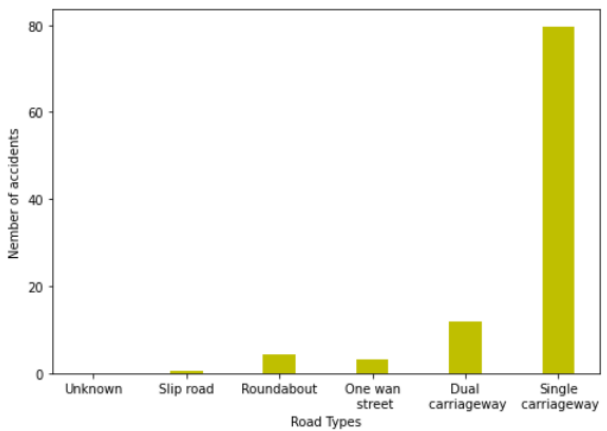
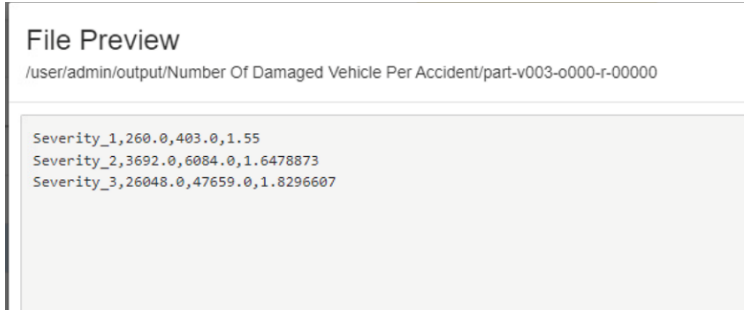


Figure 9

Q9. What is the relation between the severity and the number of cars involved in an accident?



Q9. What is the relation between the severity and the number of cars involved in an accident?

- There is a meaningful relation between the number of cars and severity of accidents.
- In severity level 3, 1.8 cars involved per accident, more than level 2 (1.7 involved cars per accidents).
- In addition, in level 2 more cars are involved than level 1 (1.6 involved cars per accidents).

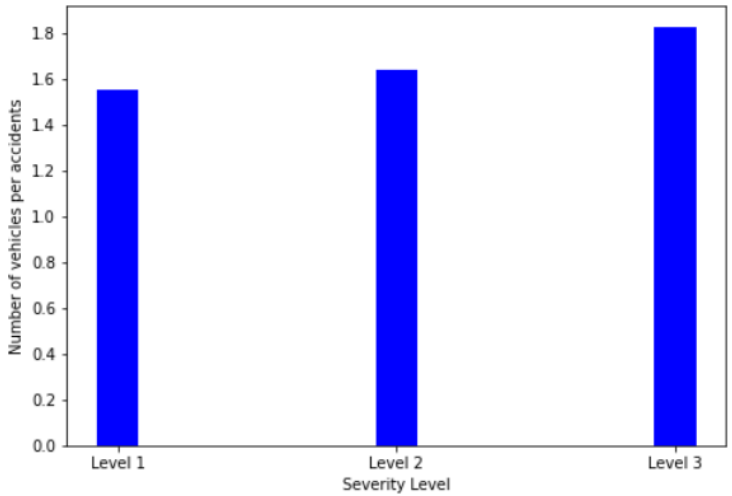
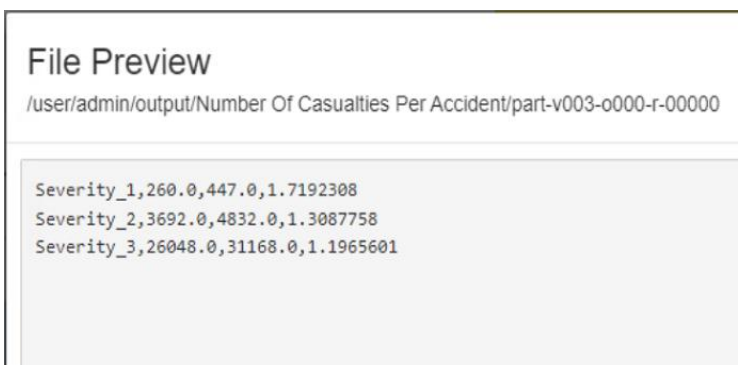
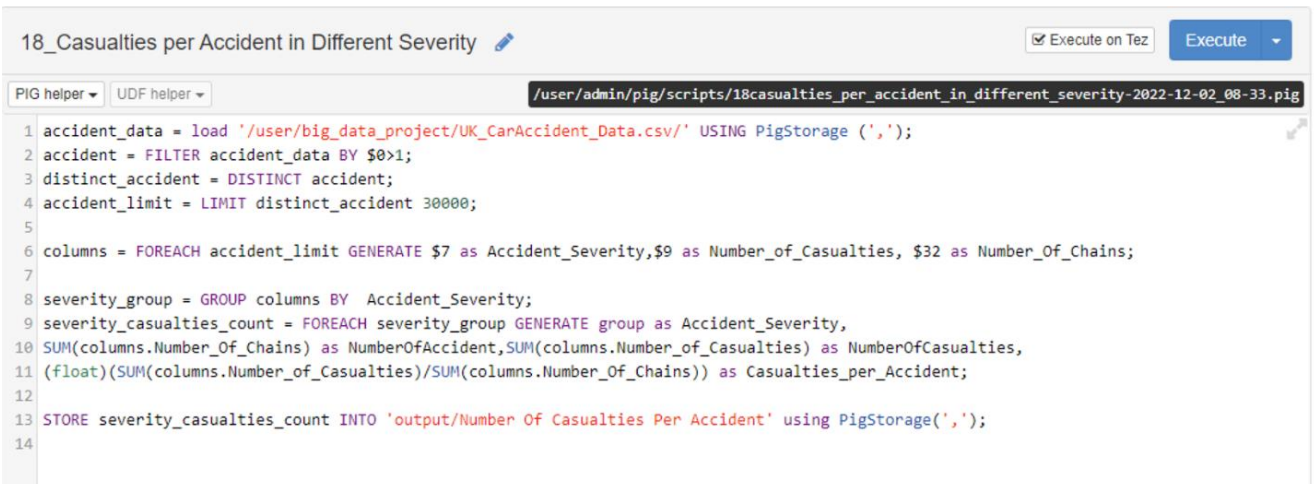


Figure 21. Number of damaged vehicles in one accident with different severity levels.

Q10. How different level of severity affect the number of injured people?





Q10. How different level of severity affect the number of injured people?

- Severity has a reverse relation with number of casualties.
- The main reason is that the total number of people in cars has a big impact on the number of casualties per accident.
- Unfortunately, this feature is not provided in our dataset.

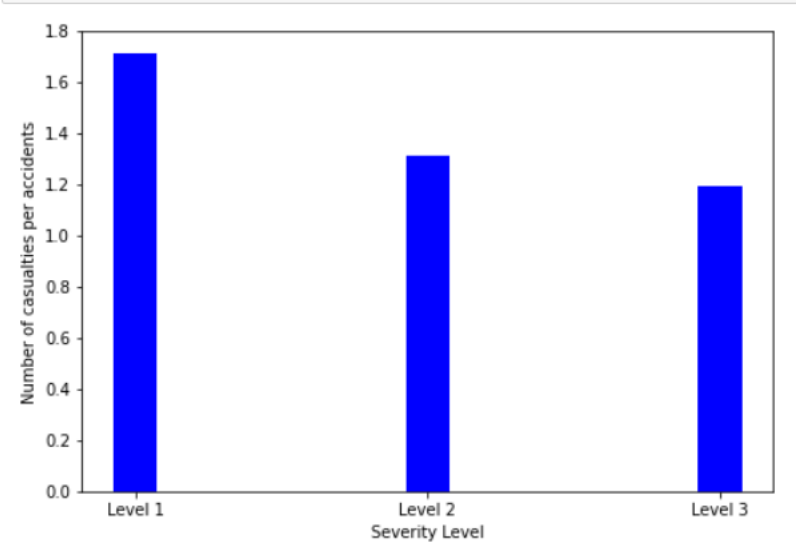


Figure 22. Number of casualties in one accident with different severities

Bibliography

- [1] Chen, C., 2017. Analysis and forecast of traffic accident big data. In *ITM Web of Conferences* (Vol. 12, p. 04029). EDP Sciences.
- [2] Dominguez-Péry, C., Tassabehji, R., Vuddaraju, L.N.R. and Duffour, V.K., 2021. Improving emergency response operations in maritime accidents using social media with big data analytics: a case study of the MV Wakashio disaster. *International Journal of Operations & Production Management*.
- [3] Jain, A.K., Kumar, A., Garg, J., Patange, U. and Jalan, P., 2015, March. TraffTrend: Real time traffic updates and traffic trends using social media analytics. In *Proceedings of the 2nd IKDD Conference on Data Sciences* (pp. 1-4).
- [4] www.kaggle.com