In [51]:
```
from pyspark.sql import SparkSession
```

In [90]:
```
from pyspark.sql.functions import col,desc
import matplotlib.pyplot as plt
import seaborn as sns
import databricks.koalas as ks
from pyspark.sql import functions as F
import plotly.express as px
import plotly.graph_objects as go
import plotly.io as pio
pio.renderers.default='notebook'
from plotly.offline import init_notebook_mode, iplot
init_notebook_mode(connected=True)
import sweetviz as sv
from pyspark.sql.functions import when
from pyspark.sql.window import Window
from pyspark.sql.functions import col, to_date, month
from pyspark.ml.feature import StringIndexer, VectorAssembler
from pyspark.ml.classification import DecisionTreeClassifier
from pyspark.ml import Pipeline
from pyspark.ml.feature import VectorIndexer
from pyspark.ml.feature import ChiSqSelector
from pyspark.ml.stat import ChiSquareTest
from pyspark.ml.stat import Correlation
from pyspark.ml.feature import VectorAssembler
from pyspark.ml.classification import LogisticRegression
from pyspark.ml import Pipeline
from pyspark.ml.evaluation import BinaryClassificationEvaluator
from pyspark.ml.classification import RandomForestClassifier
import warnings
warnings.filterwarnings("ignore")
```

In [55]:
```
spark =SparkSession.builder.appName("sample_hadoop")\
      .master("local[*]").config("spark.driver.memory","5g")\
      .config("spark.driver.host","127.0.0.1")\
      .config("spark.driver.bindAddress","127.0.0.1")\
      .config("spark.hadoop.fs.defaultFS", "hdfs://localhost:9000") \
      .getOrCreate()
```

In [56]:
```
csv_path_casualty = 'hdfs://localhost:9000/user/kiran_g/dft-road-casualty-st
csv_path_collision = "hdfs://localhost:9000/user/kiran_g/dft-road-casualty-s
csv_path_vehicle = "hdfs://localhost:9000/user/kiran_g/dft-road-casualty-sta


df_casualty = spark.read.csv(csv_path_casualty, header=True, inferSchema=Tru
df_collision = spark.read.csv(csv_path_collision, header=True, inferSchema=T
df_vehicle = spark.read.csv(csv_path_vehicle, header=True, inferSchema=True)
```

In [57]:
```
df_casualty.printSchema()
df_casualty.show(5)
```

```
root
 |-- accident_index: string (nullable = true)
 |-- accident_year: integer (nullable = true)
 |-- accident_reference: string (nullable = true)
 |-- vehicle_reference: integer (nullable = true)
 |-- casualty_reference: integer (nullable = true)
 |-- casualty_class: integer (nullable = true)
 |-- sex_of_casualty: integer (nullable = true)
 |-- age_of_casualty: integer (nullable = true)
 |-- age_band_of_casualty: integer (nullable = true)
 |-- casualty_severity: integer (nullable = true)
 |-- pedestrian_location: integer (nullable = true)
 |-- pedestrian_movement: integer (nullable = true)
 |-- car_passenger: integer (nullable = true)
 |-- bus_or_coach_passenger: integer (nullable = true)
 |-- pedestrian_road_maintenance_worker: integer (nullable = true)
 |-- casualty_type: integer (nullable = true)
 |-- casualty_home_area_type: integer (nullable = true)
 |-- casualty_imd_decile: integer (nullable = true)
 |-- lsoa_of_casualty: string (nullable = true)

+-------------+-------------+-----------------+----------------+---------
---------+--------------+--------------+--------------+------------------
-+-----------------+-------------------+-------------------+-------------+--
--------------------+-----------------------------------+-------------+------
-----------------+-------------------+----------------+
|accident_index|accident_year|accident_reference|vehicle_reference|casualty_
reference|casualty_class|sex_of_casualty|age_of_casualty|age_band_of_casualt
y|casualty_severity|pedestrian_location|pedestrian_movement|car_passenger|bu
s_or_coach_passenger|pedestrian_road_maintenance_worker|casualty_type|casual
ty_home_area_type|casualty_imd_decile|lsoa_of_casualty|
+-------------+-------------+-----------------+----------------+---------
---------+--------------+--------------+--------------+------------------
-+-----------------+-------------------+-------------------+-------------+--
--------------------+-----------------------------------+-------------+------
-----------------+-------------------+----------------+
| 197901A11AD14|         1979|         01A11AD14|               2|
1|             1|              1|             -1|                 -1|
3|                  0|                  0|           0|
0|                                  -1|          104|                     -1|
-1|             -1|
| 197901A1BAW34|         1979|         01A1BAW34|               1|
1|             3|              2|             27|                    6|
3|                 10|                  5|           0|
0|                                  -1|            0|                     -1|
-1|             -1|
| 197901A1BFD77|         1979|         01A1BFD77|               1|
1|             1|              1|             21|                    5|
3|                  0|                  0|           0|
0|                                  -1|          109|                     -1|
-1|             -1|
| 197901A1BFD77|         1979|         01A1BFD77|               1|
2|             2|              1|             20|                    4|
3|                  0|                  0|           1|
0|                                  -1|          109|                     -1|
-1|             -1|
```

```
| 197901A1BFD77|         1979|         01A1BFD77|                2|
3|             1|              1|             35|                   6|
3|                  0|                  0|           0|
0|                                  -1|          109|                     -1|
-1|              -1|
+--------------+-------------+------------------+----------------+---------
---------+--------------+---------------+---------------+-------------------
-+-----------------+------------------+------------------+-------------+--
--------------------+----------------------------------+-------------+------
-----------------+------------------+---------------+
only showing top 5 rows
```

In [58]:
```
df_collision.printSchema()
df_collision.show(5)
```

```
root
 |-- accident_index: string (nullable = true)
 |-- accident_year: integer (nullable = true)
 |-- accident_reference: string (nullable = true)
 |-- location_easting_osgr: string (nullable = true)
 |-- location_northing_osgr: string (nullable = true)
 |-- longitude: string (nullable = true)
 |-- latitude: string (nullable = true)
 |-- police_force: integer (nullable = true)
 |-- accident_severity: integer (nullable = true)
 |-- number_of_vehicles: integer (nullable = true)
 |-- number_of_casualties: integer (nullable = true)
 |-- date: string (nullable = true)
 |-- day_of_week: integer (nullable = true)
 |-- time: string (nullable = true)
 |-- local_authority_district: integer (nullable = true)
 |-- local_authority_ons_district: string (nullable = true)
 |-- local_authority_highway: string (nullable = true)
 |-- first_road_class: integer (nullable = true)
 |-- first_road_number: integer (nullable = true)
 |-- road_type: integer (nullable = true)
 |-- speed_limit: string (nullable = true)
 |-- junction_detail: integer (nullable = true)
 |-- junction_control: integer (nullable = true)
 |-- second_road_class: integer (nullable = true)
 |-- second_road_number: integer (nullable = true)
 |-- pedestrian_crossing_human_control: integer (nullable = true)
 |-- pedestrian_crossing_physical_facilities: integer (nullable = true)
 |-- light_conditions: integer (nullable = true)
 |-- weather_conditions: integer (nullable = true)
 |-- road_surface_conditions: integer (nullable = true)
 |-- special_conditions_at_site: integer (nullable = true)
 |-- carriageway_hazards: integer (nullable = true)
 |-- urban_or_rural_area: integer (nullable = true)
 |-- did_police_officer_attend_scene_of_accident: integer (nullable = true)
 |-- trunk_road_flag: integer (nullable = true)
 |-- lsoa_of_accident_location: string (nullable = true)

+--------------+-------------+------------------+----------------------+-----
-----------------+---------+--------+------------+-----------------+--------
----------+--------------------+----------+-----------+-----+--------------
---------+----------------------------+----------------------+-------------
---+-----------------+---------+-----------+---------------+----------------
+-----------------+------------------+--------------------------------+----
-----------------------------------+----------------+------------------+----
------------------+-------------------------+------------------+----------
----------+---------------------------------------------+---------------+-----
--------------------+
|accident_index|accident_year|accident_reference|location_easting_osgr|locat
ion_northing_osgr|longitude|latitude|police_force|accident_severity|number_o
f_vehicles|number_of_casualties|      date|day_of_week| time|local_authority
_district|local_authority_ons_district|local_authority_highway|first_road_cl
ass|first_road_number|road_type|speed_limit|junction_detail|junction_control
|second_road_class|second_road_number|pedestrian_crossing_human_control|pede
strian_crossing_physical_facilities|light_conditions|weather_conditions|road
_surface_conditions|special_conditions_at_site|carriageway_hazards|urban_or_
```

```
rural_area|did_police_officer_attend_scene_of_accident|trunk_road_flag|lsoa_
of_accident_location|
+-------------+-------------+-----------------+--------------------+-----
-----------------+---------+--------+-----------+----------------+--------
----------+--------------------+----------+----------+-----+--------------
---------+--------------------------+----------------------+-------------
---+-----------------+---------+----------+---------------+----------------
+-----------------+-----------------+-------------------------------+----
-----------------------------------+----------------+------------------+----
------------------+-------------------------+------------------+---------
----------+-------------------------------------------+--------------+-----
--------------------+
| 197901A11AD14|         1979|         01A11AD14|                 NULL|
NULL|     NULL|    NULL|           1|               3|                 2|
1|18/01/1979|          5|08:00|                  11|
-1|                    -1|              3|               4|        1|
30|             1|              4|               -1|                -1|
-1|                                 -1|              1|
8|                    1|                        -1|                0|
-1|                                    -1|            -1|
-1|
| 197901A1BAW34|         1979|         01A1BAW34|               198460|
894000|     NULL|    NULL|           1|               3|                1|
1|01/01/1979|          2|01:00|                  23|
-1|                    -1|              6|               0|        9|
30|             3|              4|               -1|                -1|
-1|                                 -1|              4|
8|                    3|                        -1|                0|
-1|                                    -1|            -1|
-1|
| 197901A1BFD77|         1979|         01A1BFD77|               406380|
307000|     NULL|    NULL|           1|               3|                2|
3|01/01/1979|          2|01:25|                  17|
-1|                    -1|              3|             112|        9|
30|             6|              4|               -1|                -1|
-1|                                 -1|              4|
8|                    3|                        -1|                0|
-1|                                    -1|            -1|
-1|
| 197901A1BGC20|         1979|         01A1BGC20|               281680|
440000|     NULL|    NULL|           1|               3|                2|
2|01/01/1979|          2|01:30|                   2|
-1|                    -1|              3|             502|       12|
30|             3|              2|               -1|                -1|
-1|                                 -1|              4|
8|                    3|                        -1|                0|
-1|                                    -1|            -1|
-1|
| 197901A1BGF95|         1979|         01A1BGF95|               153960|
795000|     NULL|    NULL|           1|               2|                2|
1|01/01/1979|          2|01:30|                  510|
-1|                    -1|              3|             309|        6|
30|             0|             -1|                0|                -1|
-1|                                 -1|              4|
3|                    3|                        -1|                0|
-1|                                    -1|            -1|
```

```
-1|
+-------------+-------------+------------------+---------------------+-----
----------------+---------+--------+------------+-----------------+--------
----------+--------------------+----------+-----------+-----+---------------
---------+---------------------------+----------------------+-------------
---+-----------------+---------+-----------+---------------+----------------
+-----------------+-------------------+-------------------------------+----
----------------------------------+----------------+------------------+----
-------------------+-------------------------+------------------+----------
----------+-----------------------------------------+---------------+-----
--------------------+
only showing top 5 rows
```

In [59]:
```
df_vehicle.printSchema()
df_vehicle.show(5)
```

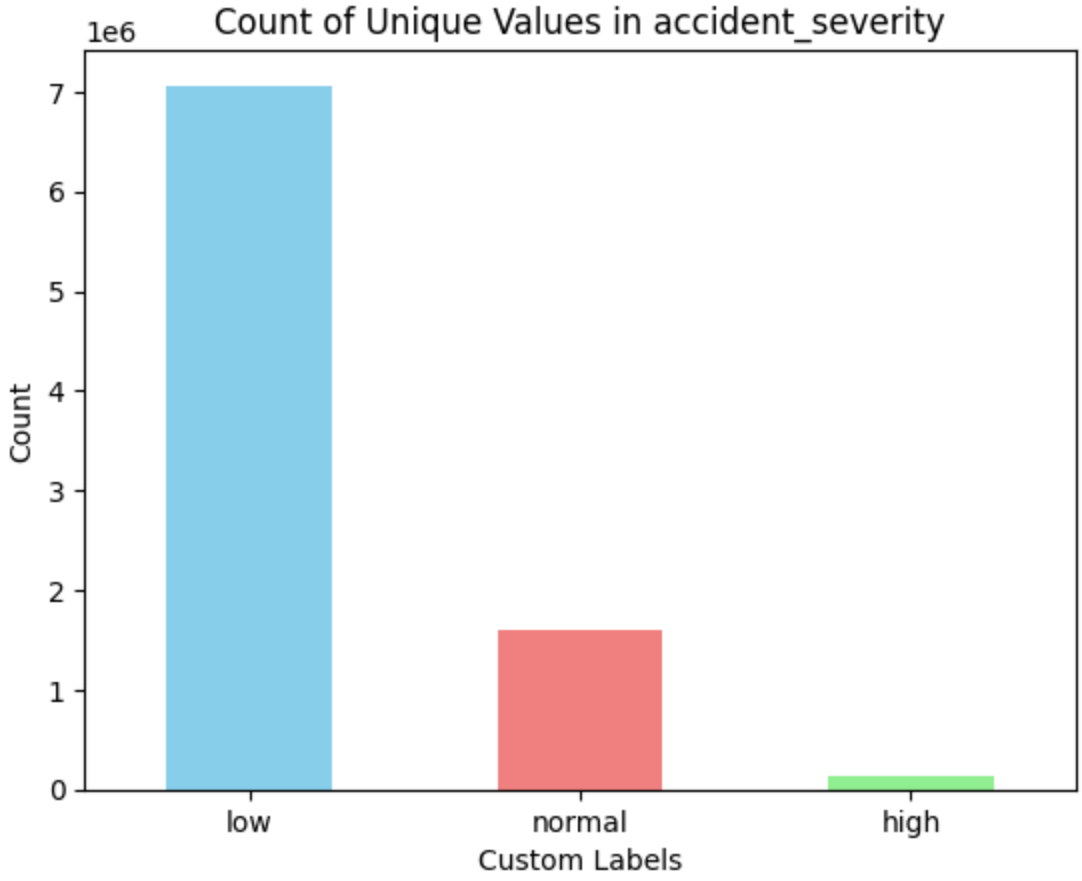
```
root
 |-- accident_index: string (nullable = true)
 |-- accident_year: integer (nullable = true)
 |-- accident_reference: string (nullable = true)
 |-- vehicle_reference: integer (nullable = true)
 |-- vehicle_type: integer (nullable = true)
 |-- towing_and_articulation: integer (nullable = true)
 |-- vehicle_manoeuvre: integer (nullable = true)
 |-- vehicle_direction_from: integer (nullable = true)
 |-- vehicle_direction_to: integer (nullable = true)
 |-- vehicle_location_restricted_lane: integer (nullable = true)
 |-- junction_location: integer (nullable = true)
 |-- skidding_and_overturning: integer (nullable = true)
 |-- hit_object_in_carriageway: integer (nullable = true)
 |-- vehicle_leaving_carriageway: integer (nullable = true)
 |-- hit_object_off_carriageway: integer (nullable = true)
 |-- first_point_of_impact: integer (nullable = true)
 |-- vehicle_left_hand_drive: integer (nullable = true)
 |-- journey_purpose_of_driver: integer (nullable = true)
 |-- sex_of_driver: integer (nullable = true)
 |-- age_of_driver: integer (nullable = true)
 |-- age_band_of_driver: integer (nullable = true)
 |-- engine_capacity_cc: integer (nullable = true)
 |-- propulsion_code: integer (nullable = true)
 |-- age_of_vehicle: integer (nullable = true)
 |-- generic_make_model: string (nullable = true)
 |-- driver_imd_decile: integer (nullable = true)
 |-- driver_home_area_type: integer (nullable = true)
 |-- lsoa_of_driver: string (nullable = true)

+--------------+-------------+------------------+-----------------+---------
---+----------------------+-----------------+----------------------+-------
-------------+---------------------------------+-----------------+-----------
-------------+-------------------------+---------------------------+--------
------------------+---------------------+----------------------+-----------
-------------+-------------+-------------+------------------+--------------
----+---------------+--------------+------------------+-----------------+---
------------------+--------------+
|accident_index|accident_year|accident_reference|vehicle_reference|vehicle_t
ype|towing_and_articulation|vehicle_manoeuvre|vehicle_direction_from|vehicle
_direction_to|vehicle_location_restricted_lane|junction_location|skidding_an
d_overturning|hit_object_in_carriageway|vehicle_leaving_carriageway|hit_obje
ct_off_carriageway|first_point_of_impact|vehicle_left_hand_drive|journey_pur
pose_of_driver|sex_of_driver|age_of_driver|age_band_of_driver|engine_capacit
y_cc|propulsion_code|age_of_vehicle|generic_make_model|driver_imd_decile|dri
ver_home_area_type|lsoa_of_driver|
+--------------+-------------+------------------+-----------------+---------
---+----------------------+-----------------+----------------------+-------
-------------+---------------------------------+-----------------+-----------
-------------+-------------------------+---------------------------+--------
------------------+---------------------+----------------------+-----------
-------------+-------------+-------------+------------------+--------------
----+---------------+--------------+------------------+-----------------+---
------------------+--------------+
| 197901A11AD14|         1979|         01A11AD14|                1|
109|                      0|               18|                     -1|
```

```
-1|                            -1|              -1|
-1|                       -1|                    -1|
-1|                  -1|                   -1|                      -1|
1|          44|                 7|                -1|             -1|
-1|                -1|              -1|                   -1|            -1
|
| 197901A11AD14|         1979|         01A11AD14|                2|
104|                      0|              13|                    -1|
-1|                            -1|              -1|
-1|                       -1|                    -1|
-1|                  -1|                   -1|                      -1|
1|          -1|                -1|                -1|             -1|
-1|                -1|              -1|                   -1|            -1
|
| 197901A1BAW34|         1979|         01A1BAW34|                1|
109|                      0|              18|                    -1|
-1|                            -1|              -1|
-1|                       -1|                    -1|
-1|                  -1|                   -1|                      -1|
1|          -1|                -1|                -1|             -1|
-1|                -1|              -1|                   -1|            -1
|
| 197901A1BFD77|         1979|         01A1BFD77|                1|
109|                      0|              18|                    -1|
-1|                            -1|              -1|
1|                       -1|                    -1|
-1|                  -1|                   -1|                      -1|
1|          21|                 5|                -1|             -1|
-1|                -1|              -1|                   -1|            -1
|
| 197901A1BFD77|         1979|         01A1BFD77|                2|
109|                      0|              18|                    -1|
-1|                            -1|              -1|
-1|                       -1|                    -1|
-1|                  -1|                   -1|                      -1|
1|          44|                 7|                -1|             -1|
-1|                -1|              -1|                   -1|            -1
|
+--------------+-------------+-----------------+-----------------+---------
---+----------------------+-----------------+----------------------+-------
-------------+------------------------------+-----------------+-----------
-------------+------------------------+--------------------------+--------
-----------------+--------------------+----------------------+-----------
--------------+-------------+-------------+-----------------+--------------
----+---------------+--------------+-----------------+-----------------+---
------------------+--------------+
only showing top 5 rows
```

In [61]:
```
# Define custom labels and colors
value_counts = df_collision.groupBy("accident_severity").count().orderBy(col
custom_labels = ['low', 'normal', 'high']# Define custom labels and colors
colors = ['skyblue', 'lightcoral', 'lightgreen']
value_counts_pd = value_counts.toPandas()# Convert Spark DataFrame to Pandas
ax = value_counts_pd.plot(kind='bar', x='accident_severity', y='count', colo
plt.xlabel('Custom Labels')
```

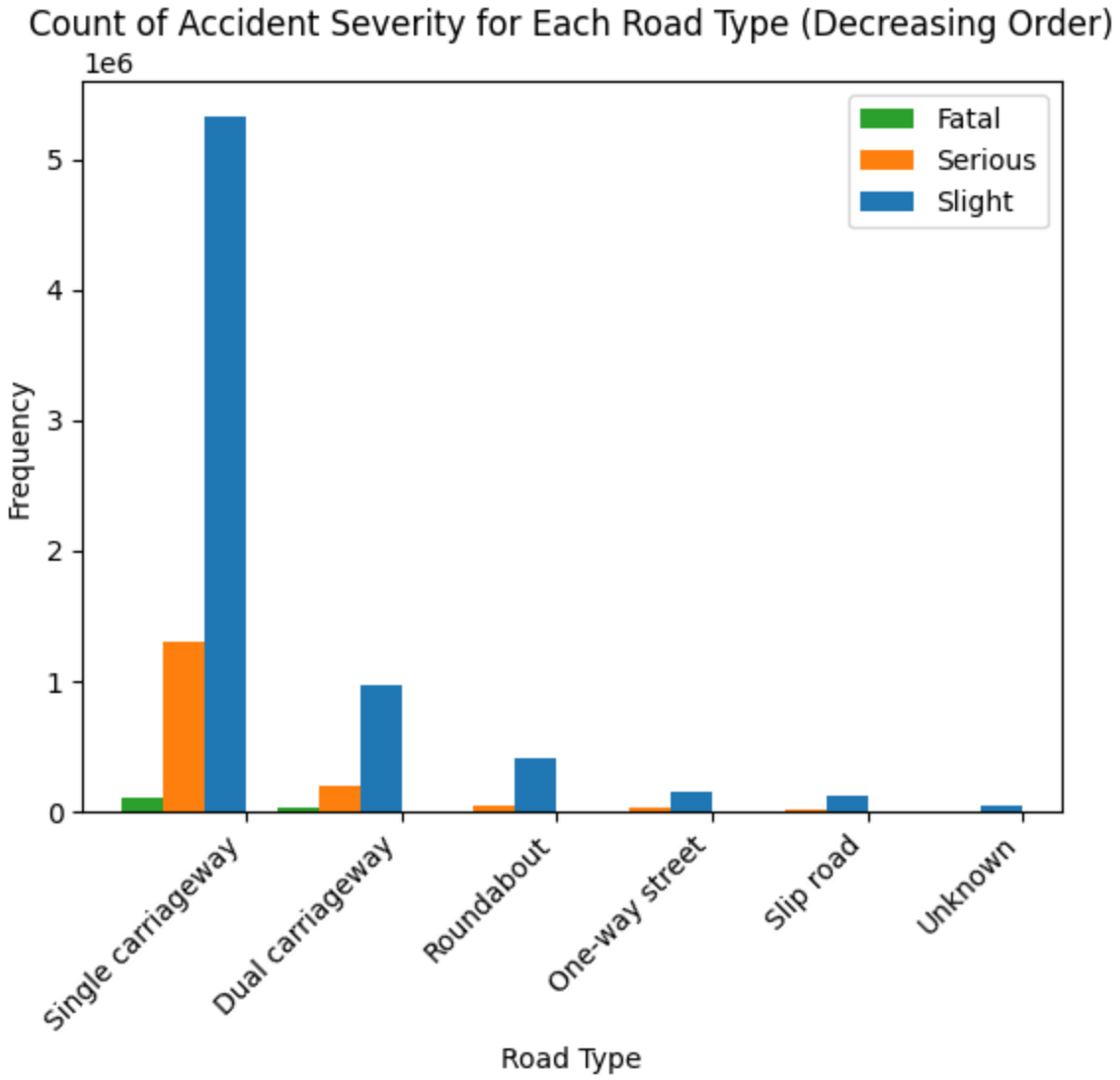
```
plt.ylabel('Count')
plt.title('Count of Unique Values in accident_severity')
ax.set_xticks(range(len(custom_labels)))
ax.set_xticklabels(custom_labels, rotation=0)
ax.get_legend().remove()
plt.show()
```



In [62]:
```
df_filtered = df_collision.filter((col("road_type") != -1) & (col("road_type
df_filtered.createOrReplaceTempView("accident_data")
result_df = spark.sql("""
    SELECT road_type, accident_severity, COUNT(*) as frequency
    FROM accident_data
    GROUP BY road_type, accident_severity
    ORDER BY road_type, accident_severity
""")
pivot_df = result_df.groupBy("road_type").pivot("accident_severity").agg(F.f
pandas_df = pivot_df.toPandas()
pandas_df['total'] = pandas_df.sum(axis=1)
pandas_df = pandas_df.sort_values(by='total', ascending=False).drop('total',
custom_colors = ['#2ca02c', '#ff7f0e', '#1f77b4']
plt.figure(figsize=(15, 7))
pandas_df.plot(kind='bar', x='road_type', stacked=False, color=custom_colors
plt.xlabel('Road Type')
plt.ylabel('Frequency')
plt.title('Count of Accident Severity for Each Road Type (Decreasing Order)'
plt.legend(['Fatal','Serious','Slight'])
ax = plt.gca()
```

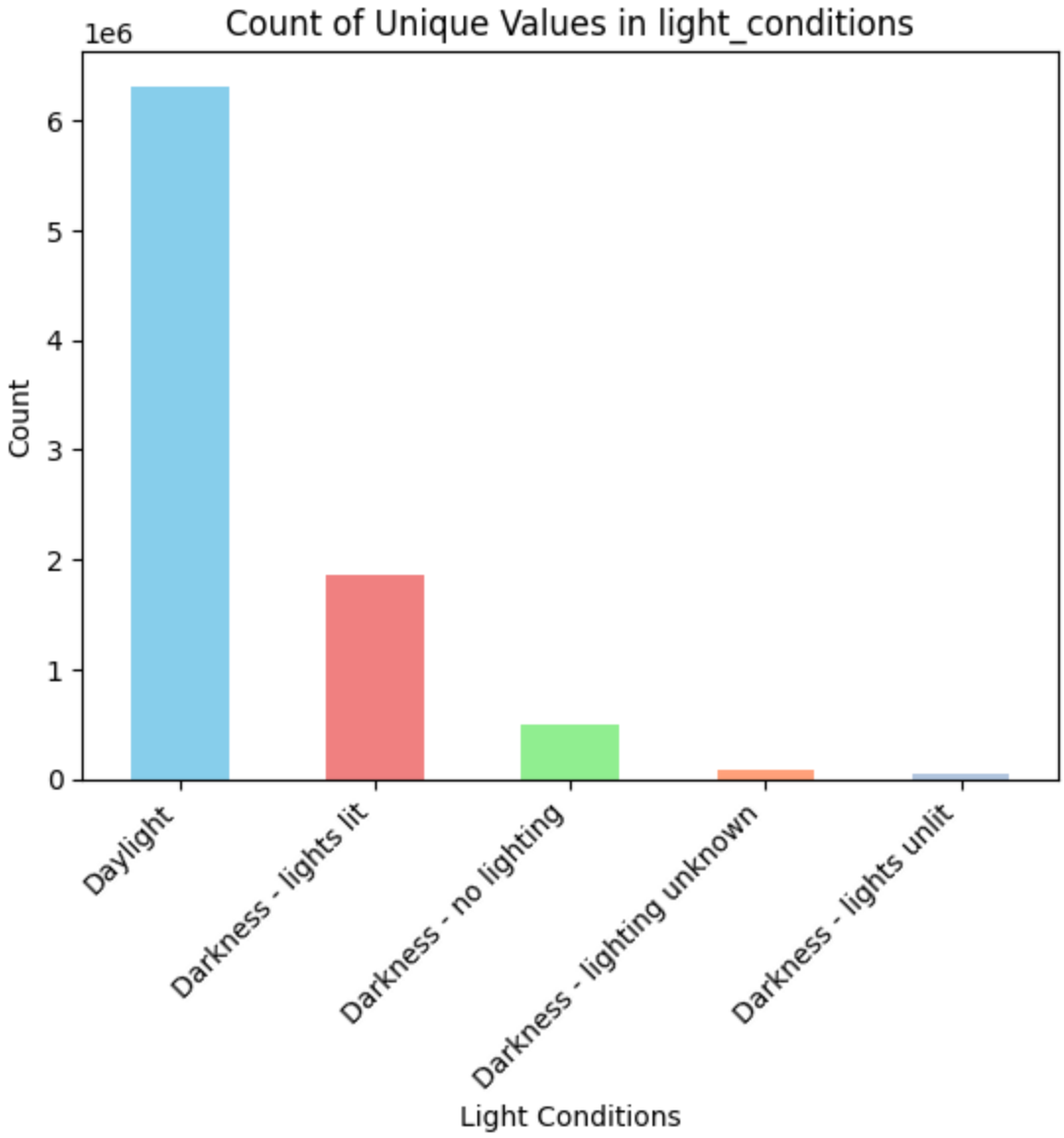
```
ax.set_xticklabels(['Single carriageway', 'Dual carriageway', 'Roundabout',
                    'One-way street',
                    'Slip road', 'Unknown'], rotation=45, ha='right')
plt.show()
```

<Figure size 1500x700 with 0 Axes>

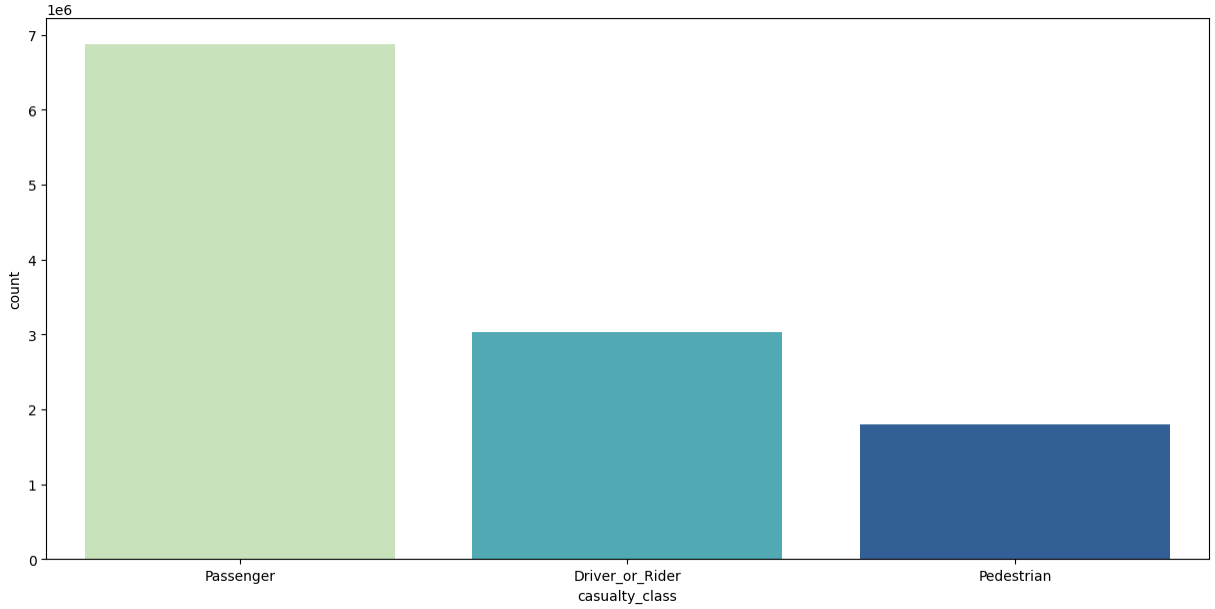


In [63]:
```
df_filtered = df_collision.filter(df_collision["light_conditions"] != -1)
value_counts = df_filtered.groupBy("light_conditions").count().orderBy("cour
pandas_value_counts = value_counts.toPandas()
custom_colors = ['skyblue', 'lightcoral', 'lightgreen', 'lightsalmon', 'ligh
ax = pandas_value_counts.plot(kind='bar', x='light_conditions', y='count', c
plt.xlabel('Light Conditions')
plt.ylabel('Count')
plt.title('Count of Unique Values in light_conditions')
ax.set_xticklabels(['Daylight',
                    'Darkness - lights lit',
                    'Darkness - no lighting',
                    'Darkness - lighting unknown',
                    'Darkness - lights unlit'], rotation=45, ha='right')
```

```
ax.get_legend().remove()
plt.show()
```



In [64]:
```
df_casualty.createOrReplaceTempView("casualty_table")
ordered_df = spark.sql("""
    SELECT casualty_class, COUNT(*) AS count
    FROM casualty_table
    GROUP BY casualty_class
    ORDER BY count DESC
""")# Run SQL query to get the ordered DataFrame
pandas_ordered_df = ordered_df.toPandas()# Convert the PySpark DataFrame to
plt.figure(figsize=(15, 7))# Create a count plot using seaborn
ax = sns.barplot(x='casualty_class', y='count', data=pandas_ordered_df, pale
                 order=pandas_ordered_df["casualty_class"])
ax.set_xticklabels(['Passenger', 'Driver_or_Rider', 'Pedestrian'])
plt.show()
```



In [65]:
```
unique_values_count = df_vehicle.select("sex_of_driver").distinct().count()
print(unique_values_count)
```

```
[Stage 713:=================================================>    (11 + 1) /
12]
4
```

In [66]:
```
unique_values_counts = df_vehicle.groupBy("sex_of_driver").count()
unique_values_counts.show()
```

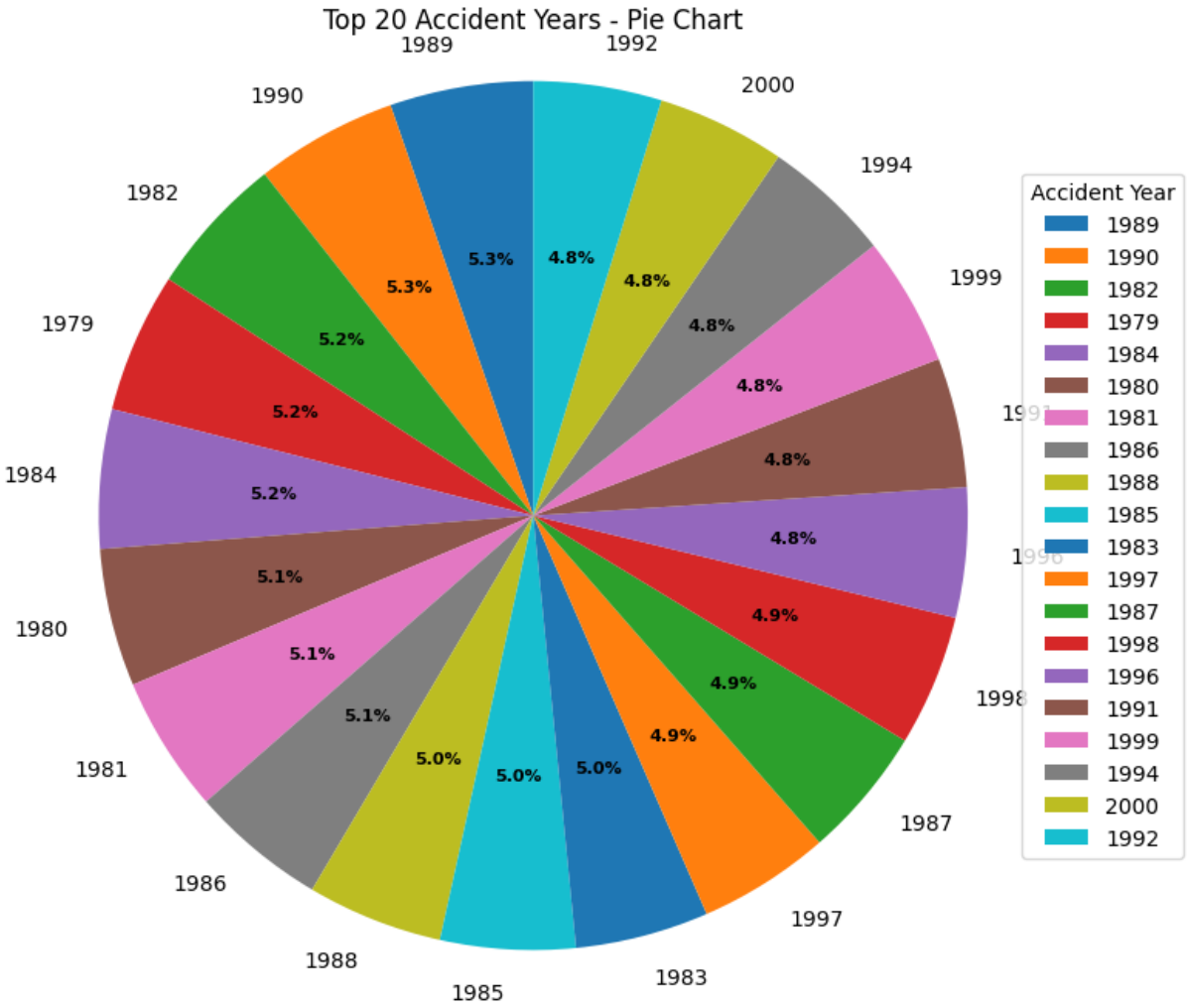
```
[Stage 719:=================================================>    (11 + 1) /
12]
+-------------+--------+
|sex_of_driver|   count|
+-------------+--------+
|           -1|   23195|
|            1|11041729|
|            3|  824160|
|            2| 3836733|
+-------------+--------+
```

In [67]:
```
df_filtered_vehicle = df_vehicle.filter(df_vehicle["sex_of_driver"] != -1)
unique_values_counts = df_filtered_vehicle.groupBy("sex_of_driver").count()
unique_values_counts.show()
```

```
[Stage 722:=================================================>    (11 + 1) /
12]
```

```
+-------------+--------+
|sex_of_driver|   count|
+-------------+--------+
|            1|11041729|
|            3|  824160|
|            2| 3836733|
+-------------+--------+
```

In [68]:
```
value_counts = df_collision.groupBy("accident_year").count().orderBy("count"
pandas_value_counts = value_counts.toPandas()# Convert PySpark DataFrame to
fig, ax = plt.subplots(figsize=(8, 8))# Convert PySpark DataFrame to Pandas
wedges, texts, autotexts = ax.pie(pandas_value_counts["count"], labels=panda
                                  autopct='%1.1f%%', startangle=90)
ax.legend(wedges, pandas_value_counts["accident_year"], title="Accident Year
plt.setp(autotexts, size=8, weight="bold")# Add legend
ax.axis("equal")# Set the aspect ratio to be equal, ensuring a circular pie
plt.title('Top 20 Accident Years - Pie Chart')
plt.show()
```



In [69]:
```
value_counts = df_collision.groupBy("accident_year").count().orderBy("count"
pandas_value_counts = value_counts.toPandas()
```

```
fig = px.pie(pandas_value_counts, names="accident_year", values="count",
             title='Top 20 Accident Years - Interactive Pie Chart',
             labels={'count': 'Accident Count', 'accident_year': 'Accident Y
             hole=0.3)
fig.update_traces(textposition='inside', textinfo='percent+label')
fig.show(renderer='browser')
```

In [70]:
```
day_mapping = {
    1: 'Monday', 2: 'Tuesday', 3: 'Wednesday', 4: 'Thursday', 5: 'Friday', 6
}
df_with_day_names = df_collision.withColumn("day_name", when(df_collision["d
                                            .when(df_collision["day_of_week"]
                                            .when(df_collision["day_of_week"]
                                            .when(df_collision["day_of_week"]
                                            .when(df_collision["day_of_week"]
                                            .when(df_collision["day_of_week"]
                                            .when(df_collision["day_of_week"]
                                            .otherwise("Unknown"))
value_counts = df_with_day_names.groupBy("day_name").count().orderBy("count"
pandas_value_counts = value_counts.toPandas()
fig = px.bar(pandas_value_counts, x="day_name", y="count",
             title='Accidents by Day of the Week',
             labels={'count': 'Accident Count', 'day_name': 'Day of the Week
             color="day_name", color_discrete_sequence=px.colors.qualitative
fig.show(renderer='browser')
```

In [71]:
```
df_with_date = df_collision.withColumn("date", to_date(col("date"), "dd/MM/y
df_with_month = df_with_date.withColumn("month", month("date"))# Extract the
value_counts = df_with_month.groupBy("month").count().orderBy("month")# Calc
pandas_value_counts = value_counts.toPandas()# Convert PySpark DataFrame to
fig = px.bar(pandas_value_counts, x="month", y="count",
             title='Accidents Count by Month',
             labels={'count': 'Accident Count', 'month': 'Month'},
             color="month", color_discrete_sequence=px.colors.qualitative.Se
fig.show(renderer='browser')
```

In [72]:
```
df_filtered_vehicle = df_collision.filter(df_collision["weather_conditions"]
unique_values_counts = df_filtered_vehicle.groupBy("weather_conditions").cou
unique_values_counts.show()# Show the result
```

```
[Stage 747:===========>                                              (2 + 8) /
10]
```

```
+------------------+-------+
|weather_conditions|  count|
+------------------+-------+
|                 1|6799093|
|                 6|  14214|
|                 3|  54158|
|                 5| 127795|
|                 9| 112614|
|                 4| 134904|
|                 8| 316249|
|                 7|  65342|
|                 2|1183420|
+------------------+-------+
```

In [73]:
```
weather_conditions_mapping = {
    1: "Rain",
    2: "Snow",
    3: "Fog",
    4: "Wind",
    5: "Ice",
    6: "Storm",
    7: "Low Sunlight",
    8: "Freezing Rain",
    9: "Hail"
}
df_with_labels = df_filtered_vehicle.withColumn("weather_conditions_label",
                                                when(df_filtered_vehicle["we
                                                .when(df_filtered_vehicle["w
                                                .when(df_filtered_vehicle["w
                                                .when(df_filtered_vehicle["w
                                                .when(df_filtered_vehicle["w
                                                .when(df_filtered_vehicle["w
                                                .when(df_filtered_vehicle["w
                                                .when(df_filtered_vehicle["w
                                                .otherwise("Unknown"))# Map
value_counts = df_with_labels.groupBy("weather_conditions_label").count()# C
pandas_value_counts = value_counts.toPandas()# Convert PySpark DataFrame to
fig = px.pie(pandas_value_counts, names="weather_conditions_label", values="
             title='Weather Conditions Distribution',
             labels={'count': 'Count', 'weather_conditions_label': 'Weather
             color="weather_conditions_label", color_discrete_sequence=px.co
fig.show(renderer='browser')
```

In [74]:
```
casualty_filtered_columns = ['accident_index', 'accident_year', 'accident_re
                             'casualty_class', 'sex_of_casualty', 'age_of_c
                             'age_band_of_casualty', 'car_passenger', 'casu
casualty_dataframe_after_cleaning = df_casualty.select(*casualty_filtered_co
casualty_dataframe_after_cleaning.select([F.sum(F.col(c).isNull().cast("int"
```

```
[Stage 753:>                                                        (0 + 10) /
10]
```

```
+--------------+-------------+-----------------+----------------+---------
-----+--------------+--------------+-----------------+-----------------
--+-------------+-------------+----------------------+
|accident_index|accident_year|accident_reference|vehicle_reference|casualty_
class|sex_of_casualty|age_of_casualty|casualty_imd_decile|age_band_of_casual
ty|car_passenger|casualty_type|casualty_home_area_type|
+--------------+-------------+-----------------+----------------+---------
-----+--------------+--------------+-----------------+-----------------
--+-------------+-------------+----------------------+
|              0|            0|                0|               0|
0|             0|             0|                0|               0|                0|
0|             0|                      0|
+--------------+-------------+-----------------+----------------+---------
-----+--------------+--------------+-----------------+-----------------
--+-------------+-------------+----------------------+
```

In [75]:
```
casualty_dataframe_after_cleaning.show(5)
```

```
+--------------+-------------+-----------------+----------------+---------
-----+--------------+--------------+-----------------+-----------------
--+-------------+-------------+----------------------+
|accident_index|accident_year|accident_reference|vehicle_reference|casualty_
class|sex_of_casualty|age_of_casualty|casualty_imd_decile|age_band_of_casual
ty|car_passenger|casualty_type|casualty_home_area_type|
+--------------+-------------+-----------------+----------------+---------
-----+--------------+--------------+-----------------+-----------------
--+-------------+-------------+----------------------+
| 197901A11AD14|         1979|         01A11AD14|               2|
1|             1|            -1|               -1|               -1|
0|           104|                      -1|
| 197901A1BAW34|         1979|         01A1BAW34|               1|
3|             2|            27|               -1|                6|
0|             0|                      -1|
| 197901A1BFD77|         1979|         01A1BFD77|               1|
1|             1|            21|               -1|                5|
0|           109|                      -1|
| 197901A1BFD77|         1979|         01A1BFD77|               1|
2|             1|            20|               -1|                4|
1|           109|                      -1|
| 197901A1BFD77|         1979|         01A1BFD77|               2|
1|             1|            35|               -1|                6|
0|           109|                      -1|
+--------------+-------------+-----------------+----------------+---------
-----+--------------+--------------+-----------------+-----------------
--+-------------+-------------+----------------------+
only showing top 5 rows
```

In [76]:
```
column_names_for_filter = ['accident_index', 'accident_year', 'accident_refe
                           'vehicle_reference', 'vehicle_type',
                           'vehicle_manoeuvre', 'vehicle_direction_from', '
                           'first_point_of_impact', 'vehicle_left_hand_driv
                           'journey_purpose_of_driver', 'sex_of_driver', 'a
                           'age_band_of_driver', 'engine_capacity_cc', 'pro
```

```
                              'age_of_vehicle', 'generic_make_model', 'driver_
                              'driver_home_area_type']
dataframe_after_cleaning_vehicle = df_vehicle.select(*column_names_for_filte
num_rows = dataframe_after_cleaning_vehicle.count()# Count the number of row
num_columns = len(dataframe_after_cleaning_vehicle.columns)# Get the number
print("Number of rows:", num_rows)# Show the results
print("Number of columns:", num_columns)
```

```
[Stage 757:=============================================>            (9 + 3) /
12]
Number of rows: 15725817
Number of columns: 20
```

In [77]:
```
num_rows = df_collision.count()
num_columns = len(df_collision.columns)
print("Number of rows:", num_rows)
print("Number of columns:", num_columns)
```

```
Number of rows: 8809915
Number of columns: 36
```

In [78]:
```
df_collision = df_collision.withColumn("longitude", df_collision["longitude"
df_collision = df_collision.withColumn("latitude", df_collision["latitude"].
df_collision_after_conversion = df_collision.dropna()
df_collision_after_conversion.show()
```

```
+--------------+-------------+-----------------+--------------------+-----
----------------+---------+---------+------------+-----------------+-------
-----------+--------------------+----------+-----------+-----+-------------
-----------+----------------------------+----------------------+------------
----+-----------------+---------+-----------+---------------+---------------
-+-----------------+-----------------+--------------------------------+---
-----------------------------------+-----------------+------------------+---
--------------------+-------------------------+------------------+--------
-----------+---------------------------------------------+---------------+----
--------------------+
|accident_index|accident_year|accident_reference|location_easting_osgr|locat
ion_northing_osgr|longitude| latitude|police_force|accident_severity|number_
of_vehicles|number_of_casualties|      date|day_of_week| time|local_authorit
y_district|local_authority_ons_district|local_authority_highway|first_road_c
lass|first_road_number|road_type|speed_limit|junction_detail|junction_contro
l|second_road_class|second_road_number|pedestrian_crossing_human_control|ped
estrian_crossing_physical_facilities|light_conditions|weather_conditions|roa
d_surface_conditions|special_conditions_at_site|carriageway_hazards|urban_or
_rural_area|did_police_officer_attend_scene_of_accident|trunk_road_flag|lsoa
_of_accident_location|
+--------------+-------------+-----------------+--------------------+-----
----------------+---------+---------+------------+-----------------+-------
-----------+--------------------+----------+-----------+-----+-------------
-----------+----------------------------+----------------------+------------
----+-----------------+---------+-----------+---------------+---------------
-+-----------------+-----------------+--------------------------------+---
-----------------------------------+-----------------+------------------+---
--------------------+-------------------------+------------------+--------
-----------+---------------------------------------------+---------------+----
--------------------+
| 1999010SU0945|         1999|         010SU0945|               519490|
203300|-0.271752| 51.71566|           1|               3|                1
|                   1|25/12/1999|          7|09:30|                     33|
E07000098|              E10000015|               1|                25|
3|         70|              5|               4|               1|
25|                                0|                                    0
|               1|                4|                      1|
0|                   0|                2|
1|              1|               E01023583|
| 1999010SU0946|         1999|         010SU0946|               521740|
201070|-0.239977|51.695137|           1|               3|                2
|                   1|17/12/1999|          6|18:38|                     33|
E07000098|              E10000015|               1|                25|
3|         70|              0|              -1|               0|
-1|                                0|                                    0
|               6|                4|                      1|
0|                   0|                2|
1|              1|               E01023584|
| 1999010SU0947|         1999|         010SU0947|               519610|
203240|-0.270037|51.715096|           1|               3|                2
|                   2|15/12/1999|          4|18:04|                     33|
E07000098|              E10000015|               1|                25|
3|         70|              0|              -1|               0|
-1|                                0|                                    0
|               4|                1|                      2|
0|                   0|                2|
```

```
1|              1|                E01023583|
| 1999010SU0948|         1999|         010SU0948|               520090|
202830|-0.263233|51.711308|           1|                2|                 2
|                    1|02/12/1999|          5|04:10|                      33|
E07000098|              E10000015|               1|                25|
3|         70|              0|             -1|                0|
-1|                                0|                                      0
|               6|                1|                          1|
0|                   0|                2|
1|              1|                E01023583|
| 1999010SU0949|         1999|         010SU0949|               522640|
200320|-0.227225|  51.6882|           1|                3|                 4
|                    3|04/12/1999|          7|09:51|                      33|
E07000098|              E10000015|               2|                1|
12|         70|              0|             -1|                0|
-1|                                0|                                      0
|               1|                1|                          2|
0|                   0|                2|
1|              1|                E01023584|
| 1999010SU0950|         1999|         010SU0950|               512390|
200290|-0.375451|51.690075|           1|                2|                 1
|                    1|29/12/1999|          4|00:30|                      33|
E07000098|              E10000015|               1|                1|
3|         70|              0|             -1|                0|
-1|                                0|                                      0
|               4|                1|                          2|
0|                   3|                2|
2|              1|                E01023528|
| 1999010SU0951|         1999|         010SU0951|               518970|
203540|-0.279194| 51.71793|           1|                3|                 2
|                    1|17/12/1999|          6|08:00|                      33|
E07000098|              E10000015|               3|             1081|
1|         70|              1|              4|                3|
1081|                               0|
0|               1|                2|                          2|
0|                   0|                2|
1|              2|                E01023583|
| 1999010SU0952|         1999|         010SU0952|               511840|
197480|-0.384303| 51.66493|           1|                2|                 1
|                    1|17/12/1999|          6|01:40|                      33|
E07000098|              E10000015|               3|             4008|
1|         40|              1|              4|                5|
0|                                0|                                       0|
4|                4|                        2|                       0|
0|                  2|                                          1|
2|               E01023554|
| 1999010SU0953|         1999|         010SU0953|               512640|
198520|-0.372406|51.674114|           1|                3|                 2
|                    1|17/11/1999|          4|08:35|                      33|
E07000098|              E10000015|               1|                1|
12|         70|              5|              4|                1|
1|                                0|                                       0|
1|                1|                        1|                       0|
0|                  2|                                         -1|
1|               E01023528|
| 1999010SU0954|         1999|         010SU0954|               521610|
```

```
201410|-0.241739| 51.69822|           1|                3|                 4
|                    1|23/11/1999|          3|17:35|                      33|
E07000098|              E10000015|               1|               25|
3|         70|              0|              -1|                0|
-1|                                  0|                                    0
|               6|                  0|               2|
0|                   0|                 2|
-1|              1|                 E01023584|
| 1999010SU0955|         1999|         010SU0955|               521630|
201350| -0.24147|51.697678|           1|                3|                 10
|                    3|26/11/1999|          6|17:00|                      33|
E07000098|              E10000015|               1|               25|
3|         70|              0|              -1|                0|
-1|                                  0|                                    0
|               6|                  0|               1|
0|                   0|                 2|
-1|              1|                 E01023584|
| 1999010SU0956|         1999|         010SU0956|               520570|
202410|-0.256434| 51.70743|           1|                3|                 2
|                    1|29/11/1999|          2|08:20|                      33|
E07000098|              E10000015|               1|               25|
3|         70|              0|              -1|                0|
-1|                                  0|                                    0
|               1|                  0|               2|
0|                   0|                 2|
-1|              1|                 E01023584|
| 1999010SU0957|         1999|         010SU0957|               518940|
203310|-0.279706| 51.71587|           1|                3|                 2
|                    1|24/11/1999|          4|18:00|                      33|
E07000098|              E10000015|               1|               25|
3|         70|              0|              -1|                0|
-1|                                  0|                                    0
|               6|                  0|               1|
0|                   0|                 2|
-1|              1|                 E01023583|
| 1999010SU0958|         1999|         010SU0958|               521550|
201510|-0.242572| 51.69913|           1|                3|                 3
|                    1|22/11/1999|          2|16:51|                      33|
E07000098|              E10000015|               1|               25|
3|         70|              0|              -1|                0|
-1|                                  0|                                    0
|               1|                  0|               1|
0|                   0|                 2|
-1|              1|                 E01023584|
| 1999010SU0959|         1999|         010SU0959|               522310|
200250|-0.232021| 51.68764|           1|                3|                 2
|                    2|12/11/1999|          6|00:48|                      33|
E07000098|              E10000015|               1|               25|
3|         70|              0|              -1|                0|
-1|                                  0|                                    0
|               4|                  0|               2|
0|                   0|                 2|
-1|              1|                 E01023584|
| 1999010SU0960|         1999|         010SU0960|               522780|
200710|-0.225064|51.691673|           1|                3|                 2
|                    2|25/11/1999|          5|08:55|                      33|
```

```
E07000098|              E10000015|               2|                1|
3|         70|              0|             -1|                0|
-1|                              0|                                   0
|               1|                1|                       2|
0|                  0|                2|
-1|              1|                E01023567|
| 1999010SU0961|         1999|         010SU0961|               522990|
202030|-0.221562| 51.70349|           1|                3|                4
|                   2|02/11/1999|          3|10:11|                      33|
E07000098|              E10000015|               2|                1|
3|         70|              0|             -1|                0|
-1|                              0|                                   0
|               1|                1|                       1|
0|                  0|                2|
-1|              1|                E01023567|
| 1999010SU0962|         1999|         010SU0962|               511880|
197440|-0.383738|51.664562|           1|                2|                2
|                   1|13/11/1999|          7|11:52|                      33|
E07000098|              E10000015|               3|             4008|
3|         40|              1|              4|                5|
0|                              0|                                    0|
1|                1|                       1|                     0|
0|                 2|                                    -1|
2|               E01023554|
| 1999010SU0963|         1999|         010SU0963|               522070|
200590|-0.235373| 51.69075|           1|                3|                4
|                   2|20/10/1999|          4|08:24|                      33|
E07000098|              E10000015|               1|               25|
3|         70|              0|             -1|                0|
-1|                              0|                                   0
|               1|                1|                       1|
0|                  0|                2|
-1|              1|                E01023584|
| 1999010SU0964|         1999|         010SU0964|               520130|
202770|-0.262675|51.710762|           1|                3|                4
|                   1|11/10/1999|          2|11:01|                      33|
E07000098|              E10000015|               1|               25|
3|         70|              0|             -1|                0|
-1|                              0|                                   0
|               1|                1|                       1|
0|                  0|                2|
-1|              1|                E01023583|
+--------------+------------+-----------------+--------------------+-----
-----------------+---------+---------+------------+-----------------+-------
-----------+-------------------+----------+----------+-----+-------------
----------+--------------------------+----------------------+------------
----+-----------------+---------+----------+---------------+--------------
-+-----------------+-----------------+-------------------------------+---
-----------------------------------+-----------------+------------------+---
--------------------+-------------------------+------------------+--------
-----------+-------------------------------------------+--------------+----
--------------------+
only showing top 20 rows
```

In [79]:
```
num_rows = df_collision_after_conversion.count()
num_columns = len(df_collision_after_conversion.columns)
print("Number of rows after removing null values:", num_rows)
print("Number of columns:", num_columns)
```

```
[Stage 766:=================================>                       (6 + 4) /
10]
Number of rows after removing null values: 3922484
Number of columns: 36
```

In [80]:
```
list_columns_after_feature_selection = ["accident_index","accident_year",'nu
df_collision_After_cleaning = df_collision_after_conversion[list_columns_aft
num_rows = df_collision_After_cleaning.count()
num_columns = len(df_collision_After_cleaning.columns)
print("Number of rows after removing null values:", num_rows)
print("Number of columns:", num_columns)
```

```
[Stage 769:=================================>                       (6 + 4) /
10]
Number of rows after removing null values: 3922484
Number of columns: 15
```

In [81]:
```
merged_df = df_collision_After_cleaning \
    .join(casualty_dataframe_after_cleaning, ['accident_index', 'accident_ye
    .join(dataframe_after_cleaning_vehicle,
          ['accident_index', 'accident_year', 'accident_reference', 'vehicle

# Show the resulting DataFrame
merged_df.show()
```

```
[Stage 787:>                                                          (0 + 2)
/ 2]
```

```
+--------------+-------------+-----------------+-----------------+---------
-----------+---------+---------+------------+----------+-----------+--------
---+---------------+-----------------+----------------+-------------------+-
----------------+-------------+--------------+--------------+------------
-------+--------------------+-------------+-------------+------------------
----+------------+-----------------+---------------------+-----------------
---+--------------------+----------------------+-----------------------+
------------+-------------+-----------------+-----------------+----------
-----+-------------+-----------------+-----------------+------------------
---+
|accident_index|accident_year|accident_reference|vehicle_reference|number_of
_casualties|longitude| latitude|police_force|     date|day_of_week|speed_li
mit|junction_detail|second_road_class|light_conditions|urban_or_rural_area|a
ccident_severity|casualty_class|sex_of_casualty|age_of_casualty|casualty_imd
_decile|age_band_of_casualty|car_passenger|casualty_type|casualty_home_area_
type|vehicle_type|vehicle_manoeuvre|vehicle_direction_from|vehicle_direction
_to|first_point_of_impact|vehicle_left_hand_drive|journey_purpose_of_driver|
sex_of_driver|age_of_driver|age_band_of_driver|engine_capacity_cc|propulsion
_code|age_of_vehicle|generic_make_model|driver_imd_decile|driver_home_area_t
ype|
+--------------+-------------+-----------------+-----------------+---------
-----------+---------+---------+------------+----------+-----------+--------
---+---------------+-----------------+----------------+-------------------+-
----------------+-------------+--------------+--------------+------------
-------+--------------------+-------------+-------------+------------------
----+------------+-----------------+---------------------+-----------------
---+--------------------+----------------------+-----------------------+
------------+-------------+-----------------+-----------------+----------
-----+-------------+-----------------+-----------------+------------------
---+
| 1999010SU0947|         1999|         010SU0947|                2|
2|-0.270037|51.715096|           1|15/12/1999|          4|         70|
0|                0|              4|                  2|                3|
2|               1|             46|                -1|                8|
1|          109|                       1|         109|               18|
4|                   8|                   2|                  -1|
-1|            1|           50|                 8|             5300|
1|            11|FORD Model Missing|               -1|                   1|
| 1999010SU0947|         1999|         010SU0947|                2|
2|-0.270037|51.715096|           1|15/12/1999|          4|         70|
0|                0|              4|                  2|                3|
2|               2|             36|                -1|                7|
2|          109|                       1|         109|               18|
4|                   8|                   2|                  -1|
-1|            1|           50|                 8|             5300|
1|            11|FORD Model Missing|               -1|                   1|
| 1999010SU0954|         1999|         010SU0954|                3|
1|-0.241739| 51.69822|           1|23/11/1999|          3|         70|
0|                0|              6|                  2|                3|
2|               2|             31|                -1|                6|
2|          109|                       3|         109|               3|
4|                   8|                   2|                  -1|
-1|            1|           35|                 6|                -1|
-1|            -1|               -1|              -1|                   3
|
| 200201KO00117|         2002|         01KO00117|                3|
```

```
3|  0.03759|51.518436|           1|13/02/2002|          4|         30|
0|                0|              1|                 1|                3|
1|              2|            43|                -1|                7|
0|          109|                      1|         109|               18|
2|                   6|                    1|                 -1|
-1|           2|          43|                7|              1388|
1|             5|       FORD FIESTA|              -1|                  1|
| 200201KO00117|         2002|         01KO00117|               3|
3|  0.03759|51.518436|           1|13/02/2002|          4|         30|
0|                0|              1|                 1|                3|
2|              1|            31|                -1|                6|
1|          109|                      1|         109|               18|
2|                   6|                    1|                 -1|
-1|           2|          43|                7|              1388|
1|             5|       FORD FIESTA|              -1|                  1|
| 2005610058305|         2005|         610058305|               2|
3|-2.678836|51.637844|          61|02/07/2005|          7|         30|
3|                5|              1|                 1|                3|
1|              1|            69|                -1|               10|
0|           9|                    -1|           9|               17|
2|                   7|                    3|                  1|
2|           1|          69|               10|                -1|
-1|           -1|               -1|              -1|                -1
|
| 2005610058705|         2005|         610058705|               3|
5| -3.04378|51.703773|          61|08/07/2005|          6|         40|
0|                0|              1|                 1|                3|
1|              2|            29|                -1|                6|
0|           9|                      1|           9|               18|
1|                   5|                    1|                  1|
15|           2|          29|                6|              1799|
1|             7|   VAUXHALL VECTRA|              -1|                  1|
| 2005610058805|         2005|         610058805|               2|
1|-3.025461|51.821545|          61|05/07/2005|          3|         30|
6|                5|              1|                 1|                3|
1|              2|            59|                -1|                9|
0|           9|                      1|           9|               10|
1|                   7|                    2|                  1|
15|           2|          59|                9|              1240|
1|             0|      NISSAN MICRA|              -1|                  1|
| 2010331007419|         2010|         331007419|               2|
4|-0.537763|52.658325|          33|17/12/2010|          6|         60|
8|                6|              6|                 2|                3|
2|              1|            66|                -1|               10|
2|          19|                      1|         19|               18|
7|                   3|                    1|                  1|
2|           1|          42|                7|                -1|
-1|           -1|               -1|              -1|                  1
|
| 2010331007419|         2010|         331007419|               2|
4|-0.537763|52.658325|          33|17/12/2010|          6|         60|
8|                6|              6|                 2|                3|
2|              2|            65|                -1|                9|
1|          19|                      1|         19|               18|
7|                   3|                    1|                  1|
2|           1|          42|                7|                -1|
```

```
-1|           -1|               -1|               -1|                    1
|
| 201601SX20709|         2016|         01SX20709|                2|
7|-0.211917| 51.58096|           1|17/06/2016|          6|         40|
6|                3|               1|                  1|                3|
1|              1|             36|                 6|                  7|
0|            9|                      1|           9|                4|
1|                   5|                    4|                    1|
6|            1|           36|                 7|               1781|
1|            25|   VOLKSWAGEN GOLF|                6|                   1|
| 201601SX20709|         2016|         01SX20709|                2|
7|-0.211917| 51.58096|           1|17/06/2016|          6|         40|
6|                3|               1|                  1|                3|
2|              1|             36|                -1|                  7|
1|            9|                     -1|           9|                4|
1|                   5|                    4|                    1|
6|            1|           36|                 7|               1781|
1|            25|   VOLKSWAGEN GOLF|                6|                   1|
| 1999010SU0945|         1999|         010SU0945|                1|
1|-0.271752| 51.71566|           1|25/12/1999|          7|         70|
5|                1|               1|                  2|                3|
1|              2|             21|                -1|                  5|
0|          109|                      1|         109|               12|
8|                   4|                    1|                   -1|
-1|           2|          21|                5|             1196|
1|            3|    VAUXHALL CORSA|               -1|                   1|
| 1999010SU0948|         1999|         010SU0948|                1|
1|-0.263233|51.711308|           1|02/12/1999|          5|         70|
0|                0|               6|                  2|                2|
1|              1|             20|                -1|                  4|
0|          109|                     -1|         109|               18|
4|                   8|                    1|                   -1|
-1|           1|          20|                4|             1275|
1|            13|    AUSTIN MAESTRO|               -1|                  -1|
| 1999010SU0949|         1999|         010SU0949|                3|
3|-0.227225|  51.6882|           1|04/12/1999|          7|         70|
0|                0|               1|                  2|                3|
1|              2|             27|                -1|                  6|
0|          109|                      1|         109|                3|
1|                   5|                    2|                   -1|
-1|           2|          27|                6|             1360|
1|            5|       PEUGEOT 306|               -1|                   1|
| 1999010SU0949|         1999|         010SU0949|                3|
3|-0.227225|  51.6882|           1|04/12/1999|          7|         70|
0|                0|               1|                  2|                3|
2|              1|             25|                -1|                  5|
1|          109|                      1|         109|                3|
1|                   5|                    2|                   -1|
-1|           2|          27|                6|             1360|
1|            5|       PEUGEOT 306|               -1|                   1|
| 1999010SU0958|         1999|         010SU0958|                1|
1|-0.242572| 51.69913|           1|22/11/1999|          2|         70|
0|                0|               1|                  2|                3|
1|              1|             53|                -1|                  8|
0|          109|                      1|         109|               18|
4|                   8|                    1|                   -1|
```

```
-1|           1|          53|                8|               -1|
-1|           -1|              -1|              -1|                    1
|
| 200201KO00107|         2002|         01KO00107|                1|
2| 0.032337|   51.517|           1|08/01/2002|          3|         30|
6|                3|              1|                  1|               3|
2|              2|             38|                 -1|                  7|
1|          109|                      1|         109|                9|
7|                   5|                    4|                      -1|
-1|           2|          65|                 9|              1275|
1|            3|      NISSAN MICRA|              -1|                  1|
| 200201KO00120|         2002|         01KO00120|                1|
1| 0.042837|51.558365|           1|14/02/2002|          5|         30|
0|                0|              4|                  1|               3|
2|              2|             18|                 -1|                  4|
2|          109|                      1|         109|                2|
0|                   0|                    2|                      -1|
-1|           1|          17|                 4|              1117|
1|           11|       FORD FIESTA|              -1|                  1|
| 200201KO00125|         2002|         01KO00125|                1|
1| 0.054774| 51.55114|           1|26/01/2002|          7|         30|
0|                0|              4|                  1|               3|
3|              2|             42|                 -1|                  7|
0|           0|                      1|         109|               18|
4|                   8|                    1|                      -1|
-1|           3|          -1|               -1|               -1|
-1|           -1|              -1|              -1|                   -1
|
+-------------+------------+-----------------+----------------+---------
----------+---------+---------+-----------+----------+----------+--------
---+---------------+-----------------+----------------+------------------+-
---------------+-------------+--------------+--------------+-----------
-------+--------------------+-------------+-------------+------------------
----+------------+-----------------+---------------------+-----------------
---+--------------------+----------------------+-----------------------+
------------+-------------+-----------------+----------------+----------
-----+-------------+-----------------+----------------+-----------------
---+
only showing top 20 rows
```

In [82]:
```
num_rows = merged_df.count()
num_columns = len(merged_df.columns)
print("Number of rows after removing null values:", num_rows)
print("Number of columns:", num_columns)
```

```
[Stage 798:==============================>                          (6 + 5) /
11]
Number of rows after removing null values: 5270471
Number of columns: 40
```

In [83]:
```
df_complete_after_removing_null_values = merged_df.dropna()
num_rows = df_complete_after_removing_null_values.count()# Count the number
num_columns = len(df_complete_after_removing_null_values.columns)# Get the r
```

```
print("Number of rows after removing null values:", num_rows)# Show the resu
print("Number of columns:", num_columns)
```

```
[Stage 815:=================================================>    (21 + 2) /
23]
Number of rows after removing null values: 5270471
Number of columns: 40
```

In [84]:
```
string_columns = [c for c, t in df_complete_after_removing_null_values.dtype
complete_dataset_after_removing_object = df_complete_after_removing_null_val
complete_dataset_after_removing_object.show()# Show the resulting DataFrame
```

```
[Stage 827:=============================================>          (9 + 2) /
11]
```

```
+-------------+-----------------+--------------------+---------+---------+--
----------+-----------+--------------+-----------------+---------------+--
----------------+-----------------+-------------+--------------+---------
------+--------------------+--------------------+-------------+-------------+
----------------------+------------+-----------------+--------------------
-+--------------------+---------------------+----------------------+-------
-----------------+-------------+-------------+------------------+----------
--------+--------------+-------------+----------------+-----------------
---+
|accident_year|vehicle_reference|number_of_casualties|longitude| latitude|po
lice_force|day_of_week|junction_detail|second_road_class|light_conditions|ur
ban_or_rural_area|accident_severity|casualty_class|sex_of_casualty|age_of_ca
sualty|casualty_imd_decile|age_band_of_casualty|car_passenger|casualty_type|
casualty_home_area_type|vehicle_type|vehicle_manoeuvre|vehicle_direction_fro
m|vehicle_direction_to|first_point_of_impact|vehicle_left_hand_drive|journey
_purpose_of_driver|sex_of_driver|age_of_driver|age_band_of_driver|engine_cap
acity_cc|propulsion_code|age_of_vehicle|driver_imd_decile|driver_home_area_t
ype|
+-------------+-----------------+--------------------+---------+---------+--
----------+-----------+--------------+-----------------+---------------+--
----------------+-----------------+-------------+--------------+---------
------+--------------------+--------------------+-------------+-------------+
----------------------+------------+-----------------+--------------------
-+--------------------+---------------------+----------------------+-------
-----------------+-------------+-------------+------------------+----------
--------+--------------+-------------+----------------+-----------------
---+
|         1999|                2|                   2|-0.270037|51.715096|
1|          4|             0|                0|             4|
2|                3|             2|              1|             46|
-1|                   8|            1|          109|                      1|
109|               18|                     4|                   8|
2|                   -1|                      -1|            1|
50|                 8|              5300|              1|            11|
-1|                   1|
|         1999|                2|                   2|-0.270037|51.715096|
1|          4|             0|                0|             4|
2|                3|             2|              2|             36|
-1|                   7|            2|          109|                      1|
109|               18|                     4|                   8|
2|                   -1|                      -1|            1|
50|                 8|              5300|              1|            11|
-1|                   1|
|         1999|                1|                   1|-0.236371|51.691128|
1|          7|             0|                0|             1|
2|                3|             1|              1|             63|
-1|                   9|            0|          109|                     -1|
109|               18|                     4|                   8|
1|                   -1|                      -1|            1|
63|                 9|               -1|            -1|            -1|
-1|                  -1|
|         1999|                1|                   1|  0.22863|51.628544|
1|          5|             0|                0|             4|
2|                3|             1|              1|             24|
-1|                   5|            0|          109|                     -1|
109|               18|                     8|                   4|
```

```
1|                  -1|                    -1|           1|
24|                5|             1850|             1|            0|
-1|                -1|
|         1999|               1|                  1|-0.167326|51.501766|
1|          6|              0|                0|              1|
1|                3|             3|              2|             22|
-1|                5|            0|             0|                      1|
109|              18|                     7|                   3|
-1|                   -1|                    -1|            1|
-1|               -1|               -1|            -1|            -1|
-1|                -1|
|         1999|               1|                  1|-0.153377|51.501095|
1|          2|              0|                0|              1|
1|                3|             1|              1|             22|
-1|                5|            0|             3|                      1|
3|               18|                   6|                  2|
4|                -1|                    -1|            1|
22|                5|               85|             1|            4|
-1|                1|
|         1999|               1|                  1|-0.147661|51.492825|
1|          6|              0|                0|              1|
1|                3|             1|              1|             34|
-1|                6|            0|             11|                      1|
11|               18|                   6|                  2|
-1|                -1|                    -1|            1|
34|                6|               -1|             2|            8|
-1|                1|
|         1999|               1|                  1|-0.144683|  51.4881|
1|          4|              6|                6|              1|
1|                3|             3|              2|              2|
-1|                1|            0|             0|                     -1|
109|              1|                     8|                   2|
2|                -1|                    -1|            1|
56|                9|               -1|            -1|            -1|
-1|                1|
|         1999|               1|                  1|-0.129526| 51.49946|
1|          6|              3|                5|              1|
1|                3|             1|              1|             43|
-1|                7|            0|           106|                     -1|
106|              18|                     2|                   6|
1|                -1|                    -1|            1|
43|                7|              248|             1|           18|
-1|                -1|
|         1999|               1|                  1|-0.160267|51.501743|
1|          6|              7|                3|              1|
1|                2|             3|              2|             18|
-1|                4|            0|             0|                      1|
109|              18|                     2|                   6|
4|                -1|                    -1|            3|
-1|               -1|               -1|            -1|            -1|
-1|                -1|
|         1999|               2|                  1| -0.14377| 51.49285|
1|          6|              3|                3|              1|
1|                3|             1|              2|             31|
-1|                6|            0|             1|                      1|
1|               18|                   6|                  2|
```

```
3|                  -1|                      -1|           2|
31|                6|               -1|           -1|           -1|
-1|                  1|
|        1999|               2|                  1|-0.159244|51.502087|
1|         7|              3|                6|              1|
1|               3|            1|              1|             -1|
-1|                 -1|           0|          109|                      1|
109|               4|                    6|                  2|
2|                  -1|                      -1|           1|
-1|               -1|              1587|             1|           13|
-1|                  1|
|        1999|               1|                  1| -0.13239| 51.49285|
1|         3|              6|                6|              4|
1|               3|            1|              1|             -1|
-1|                 -1|           0|          109|                      1|
109|              18|                    5|                  1|
4|                  -1|                      -1|           1|
-1|               -1|              1295|             1|           12|
-1|                  1|
|        1999|               1|                  1|-0.143218| 51.49581|
1|         4|              0|                0|              1|
1|               3|            2|             2|             70|
-1|                 10|           0|          11|                     -1|
11|               3|                  4|                  8|
0|                  -1|                      -1|           1|
56|                9|              9600|             2|           11|
-1|                  1|
|        1999|               2|                  1|-0.135517|51.497395|
1|         5|              6|                4|              1|
1|               3|            1|              1|             21|
-1|                  5|           0|          106|                      1|
106|              18|                    2|                  6|
1|                  -1|                      -1|           1|
21|                5|              400|             1|           11|
-1|                  1|
|        1999|               1|                  1|-0.123043|51.488922|
1|         7|              0|                0|              1|
1|               3|            1|              1|             55|
-1|                  8|           0|          109|                     -1|
109|               3|                    6|                  2|
2|                  -1|                      -1|           1|
55|                8|               -1|           -1|           -1|
-1|                 -1|
|        1999|               1|                  1|-0.125686| 51.49472|
1|         6|              1|                3|              1|
1|               3|            1|              1|             34|
-1|                  6|           0|          109|                     -1|
109|               3|                    7|                  3|
2|                  -1|                      -1|           1|
34|                6|               -1|           -1|           -1|
-1|                 -1|
|        1999|               1|                  1|-0.148596|51.494637|
1|         6|              6|                5|              1|
1|               3|            1|             2|             26|
-1|                  6|           0|           3|                      1|
3|                18|                   6|                2|
```

```
1|                    -1|                        -1|            2|
26|                  6|                 85|              1|             9|
-1|                     1|
|         1999|                  1|                    1|-0.139034|51.492416|
1|           6|                3|                 6|                1|
1|                 3|              3|              1|              37|
-1|                    7|             0|             0|                       -1|
3|                18|                      2|                    6|
1|                     -1|                        -1|             1|
24|                  5|                124|              1|             4|
-1|                     1|
|         1999|                  1|                    1|-0.143927| 51.49609|
1|           3|                0|                 0|                4|
1|                 3|              3|              1|              18|
-1|                    4|             0|             0|                       1|
108|                18|                      7|                    3|
-1|                     -1|                        -1|             1|
52|                  8|               2664|              2|             7|
-1|                     1|
+-------------+-----------------+--------------------+---------+---------+--
----------+-----------+---------------+-----------------+----------------+--
----------------+-----------------+-------------+--------------+---------
------+--------------------+--------------------+-------------+-------------+
----------------------+------------+-----------------+--------------------
-+--------------------+--------------------+----------------------+-------
------------------+-------------+-------------+------------------+----------
--------+---------------+-------------+-----------------+-----------------
---+
only showing top 20 rows
```

In [85]:
```
num_rows = complete_dataset_after_removing_object.count()
num_columns = len(complete_dataset_after_removing_object.columns)
print("Number of rows after removing null values:", num_rows)
print("Number of columns:", num_columns)
```

```
[Stage 843:=================================================>    (21 + 2) /
23]
Number of rows after removing null values: 5270471
Number of columns: 35
```

In [86]:
```
complete_dataset_after_removing_object.printSchema()
```

```
root
 |-- accident_year: integer (nullable = true)
 |-- vehicle_reference: integer (nullable = true)
 |-- number_of_casualties: integer (nullable = true)
 |-- longitude: float (nullable = true)
 |-- latitude: float (nullable = true)
 |-- police_force: integer (nullable = true)
 |-- day_of_week: integer (nullable = true)
 |-- junction_detail: integer (nullable = true)
 |-- second_road_class: integer (nullable = true)
 |-- light_conditions: integer (nullable = true)
 |-- urban_or_rural_area: integer (nullable = true)
 |-- accident_severity: integer (nullable = true)
 |-- casualty_class: integer (nullable = true)
 |-- sex_of_casualty: integer (nullable = true)
 |-- age_of_casualty: integer (nullable = true)
 |-- casualty_imd_decile: integer (nullable = true)
 |-- age_band_of_casualty: integer (nullable = true)
 |-- car_passenger: integer (nullable = true)
 |-- casualty_type: integer (nullable = true)
 |-- casualty_home_area_type: integer (nullable = true)
 |-- vehicle_type: integer (nullable = true)
 |-- vehicle_manoeuvre: integer (nullable = true)
 |-- vehicle_direction_from: integer (nullable = true)
 |-- vehicle_direction_to: integer (nullable = true)
 |-- first_point_of_impact: integer (nullable = true)
 |-- vehicle_left_hand_drive: integer (nullable = true)
 |-- journey_purpose_of_driver: integer (nullable = true)
 |-- sex_of_driver: integer (nullable = true)
 |-- age_of_driver: integer (nullable = true)
 |-- age_band_of_driver: integer (nullable = true)
 |-- engine_capacity_cc: integer (nullable = true)
 |-- propulsion_code: integer (nullable = true)
 |-- age_of_vehicle: integer (nullable = true)
 |-- driver_imd_decile: integer (nullable = true)
 |-- driver_home_area_type: integer (nullable = true)
```

In [ ]:

In [87]:
```
x = ["urban_or_rural_area", "light_conditions", "day_of_week", "casualty_typ
y = "accident_severity"
if "features" in complete_dataset_after_removing_object.columns:
    complete_dataset_after_removing_object = complete_dataset_after_removing
assembler = VectorAssembler(inputCols=x, outputCol="features_temp")# Create
assembled_df = assembler.transform(complete_dataset_after_removing_object)
assembled_df = assembled_df.drop("features_temp")
(train_data, test_data) = assembled_df.randomSplit([0.8, 0.2], seed=42)
assembled_df = assembled_df.drop("features_temp")
lr = LogisticRegression(featuresCol="features_temp", labelCol=y)
pipeline = Pipeline(stages=[assembler, lr])
model = pipeline.fit(train_data)
```

```
[Stage 1137:=========================================>              (8 + 3) /
11]
```

In [88]:
```
predictions = model.transform(test_data)
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
evaluator = MulticlassClassificationEvaluator(predictionCol="prediction", la
accuracy = evaluator.evaluate(predictions)
print("Accuracy:", accuracy)
```

```
[Stage 1148:=====================================>                 (16 + 7) /
23]
Accuracy: 0.8202823451253944
```

In [89]:
```
rf = RandomForestClassifier(featuresCol="features_temp", labelCol=y, numTree
pipeline = Pipeline(stages=[assembler, rf])# Create a pipeline with the Vect
model = pipeline.fit(train_data)# Train the model
predictions = model.transform(test_data)# Make predictions on the test set
evaluator = MulticlassClassificationEvaluator(predictionCol="prediction", la
accuracy = evaluator.evaluate(predictions)
print("Random Forest Accuracy:", accuracy)
```

```
[Stage 1241:=====================================>                 (16 + 7) /
23]
Random Forest Accuracy: 0.8202946828955797
```

In [94]:
```
dt = DecisionTreeClassifier(featuresCol="features_temp", labelCol=y, maxDept
pipeline = Pipeline(stages=[assembler, dt])# Create a pipeline with the Vect
model = pipeline.fit(train_data)# Train the model
predictions = model.transform(test_data)# Make predictions on the test set
evaluator = MulticlassClassificationEvaluator(predictionCol="prediction", la
accuracy = evaluator.evaluate(predictions)
print("Decision Tree Accuracy:", accuracy)
```

```
[Stage 1370:================================================>      (21 + 2) /
23]
Decision Tree Accuracy: 0.8210415925214132
```

In [ ]:

In [ ]:

In [ ]: