PRODIGY_DS_05

[1]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

[2]:

```
df=pd.read_csv("/content/RTA Dataset.csv")
df.head()
```

[2] :

```
       Time Day_of_week Age_band_of_driver Sex_of_driver      Educational_level  \
0  17:02:00      Monday              18-30          Male    Above high school
1  17:02:00      Monday              31-50          Male   Junior high school
2  17:02:00      Monday              18-30          Male   Junior high school
3   1:06:00      Sunday              18-30          Male   Junior high school
4   1:06:00      Sunday              18-30          Male   Junior high school

  Vehicle_driver_relation Driving_experience      Type_of_vehicle  \
0                Employee              1-2yr           Automobile
1                Employee          Above 10yr  Public (> 45 seats)
2                Employee              1-2yr       Lorry (41?100Q)
3                Employee             5-10yr  Public (> 45 seats)
4                Employee              2-5yr                  NaN

  Owner_of_vehicle Service_year_of_vehicle  ... Vehicle_movement  \
0            Owner              Above 10yr  ...  Going straight
1            Owner                 5-10yrs  ...  Going straight
2            Owner                     NaN  ...  Going straight
3     Governmental                     NaN  ...  Going straight
4            Owner                 5-10yrs  ...  Going straight

    Casualty_class Sex_of_casualty Age_band_of_casualty Casualty_severity  \
0               na              na                   na                na
1               na              na                   na                na
2  Driver or rider            Male                31-50                 3
3       Pedestrian          Female                18-30                 3
4               na              na                   na                na
```

```
  Work_of_casuality  Fitness_of_casuality  Pedestrian_movement  \
0               NaN                   NaN     Not a Pedestrian
1               NaN                   NaN     Not a Pedestrian
2            Driver                   NaN     Not a Pedestrian
3            Driver                Normal     Not a Pedestrian
4               NaN                   NaN     Not a Pedestrian

            Cause_of_accident Accident_severity
0             Moving Backward    Slight Injury
1                  Overtaking    Slight Injury
2   Changing lane to the left   Serious Injury
3  Changing lane to the right    Slight Injury
4                  Overtaking    Slight Injury

[5 rows x 32 columns]
```

[3]:

```
df.shape
```

[3]:

```
(5993, 32)
```

[4]:

```
df.describe()
```

[4]:

```
       Number_of_vehicles_involved  Number_of_casualties
count                  5992.000000           5992.000000
mean                      1.972964              1.465621
std                       0.624651              0.928860
min                       1.000000              1.000000
25%                       2.000000              1.000000
50%                       2.000000              1.000000
75%                       2.000000              2.000000
max                       6.000000              8.000000
```

[5]:

```
df.describe(include="all")
```

[5]:

```
            Time Day_of_week Age_band_of_driver Sex_of_driver  \
count       5993        5993               5993          5992
unique       927           7                  6             3
top     16:00:00      Friday              31-50          Male
freq          57         975               2042          5483
mean         NaN         NaN                NaN           NaN
std          NaN         NaN                NaN           NaN
min          NaN         NaN                NaN           NaN
25%          NaN         NaN                NaN           NaN
50%          NaN         NaN                NaN           NaN
75%          NaN         NaN                NaN           NaN
max          NaN         NaN                NaN           NaN
```

```
         Educational_level  Vehicle_driver_relation  Driving_experience   \
count                 5643                     5746                5595
unique                   7                        4                   7
top     Junior high school                 Employee              5-10yr
freq                  3696                     4630                1665
mean                   NaN                      NaN                 NaN
std                    NaN                      NaN                 NaN
min                    NaN                      NaN                 NaN
25%                    NaN                      NaN                 NaN
50%                    NaN                      NaN                 NaN
75%                    NaN                      NaN                 NaN
max                    NaN                      NaN                 NaN

       Type_of_vehicle  Owner_of_vehicle  Service_year_of_vehicle   ...  \
count             5511              5762                     4019  ...
unique              17                 4                        6  ...
top         Automobile             Owner                  Unknown  ...
freq              1573              5088                     1377  ...
mean               NaN               NaN                      NaN  ...
std                NaN               NaN                      NaN  ...
min                NaN               NaN                      NaN  ...
25%                NaN               NaN                      NaN  ...
50%                NaN               NaN                      NaN  ...
75%                NaN               NaN                      NaN  ...
max                NaN               NaN                      NaN  ...

       Vehicle_movement    Casualty_class  Sex_of_casualty  Age_band_of_casualty  \
count              5870              5992             5992                  5992
unique               13                 4                3                     6
top      Going straight  Driver or rider              Male                    na
freq               4033              2344             2507                  2105
mean                NaN               NaN              NaN                   NaN
std                 NaN               NaN              NaN                   NaN
min                 NaN               NaN              NaN                   NaN
25%                 NaN               NaN              NaN                   NaN
50%                 NaN               NaN              NaN                   NaN
75%                 NaN               NaN              NaN                   NaN
max                 NaN               NaN              NaN                   NaN

       Casualty_severity  Work_of_casuality  Fitness_of_casuality   \
count               5992               4430                  4692
unique                 4                  7                     5
top                    3             Driver                Normal
freq                3395               2862                  4656
mean                 NaN                NaN                   NaN
std                  NaN                NaN                   NaN
```

```
min                     NaN               NaN                   NaN
25%                     NaN               NaN                   NaN
50%                     NaN               NaN                   NaN
75%                     NaN               NaN                   NaN
max                     NaN               NaN                   NaN

       Pedestrian_movement Cause_of_accident Accident_severity
count                 5992              5992              5992
unique                   9                20                 3
top       Not a Pedestrian     No distancing     Slight Injury
freq                  5523              1104              5177
mean                   NaN               NaN               NaN
std                    NaN               NaN               NaN
min                    NaN               NaN               NaN
25%                    NaN               NaN               NaN
50%                    NaN               NaN               NaN
75%                    NaN               NaN               NaN
max                    NaN               NaN               NaN

[11 rows x 32 columns]
```

[6]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5993 entries, 0 to 5992
Data columns (total 32 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Time                         5993 non-null   object
 1   Day_of_week                  5993 non-null   object
 2   Age_band_of_driver           5993 non-null   object
 3   Sex_of_driver                5992 non-null   object
 4   Educational_level            5643 non-null   object
 5   Vehicle_driver_relation      5746 non-null   object
 6   Driving_experience           5595 non-null   object
 7   Type_of_vehicle              5511 non-null   object
 8   Owner_of_vehicle             5762 non-null   object
 9   Service_year_of_vehicle      4019 non-null   object
 10  Defect_of_vehicle            3791 non-null   object
 11  Area_accident_occured        5874 non-null   object
 12  Lanes_or_Medians             5793 non-null   object
 13  Road_allignment              5925 non-null   object
 14  Types_of_Junction            5992 non-null   object
 15  Road_surface_type            5911 non-null   object
 16  Road_surface_conditions      5992 non-null   object
 17  Light_conditions             5992 non-null   object
 18  Weather_conditions           5992 non-null   object
 19  Type_of_collision            5927 non-null   object
 20  Number_of_vehicles_involved  5992 non-null   float64
 21  Number_of_casualties         5992 non-null   float64
 22  Vehicle_movement             5870 non-null   object
 23  Casualty_class               5992 non-null   object
 24  Sex_of_casualty              5992 non-null   object
 25  Age_band_of_casualty         5992 non-null   object
 26  Casualty_severity            5992 non-null   object
 27  Work_of_casuality            4430 non-null   object
 28  Fitness_of_casuality         4692 non-null   object
 29  Pedestrian_movement          5992 non-null   object
 30  Cause_of_accident            5992 non-null   object
 31  Accident_severity            5992 non-null   object
dtypes: float64(2), object(30)
memory usage: 1.5+ MB
```

[7] :
```
df.duplicated().sum()
```

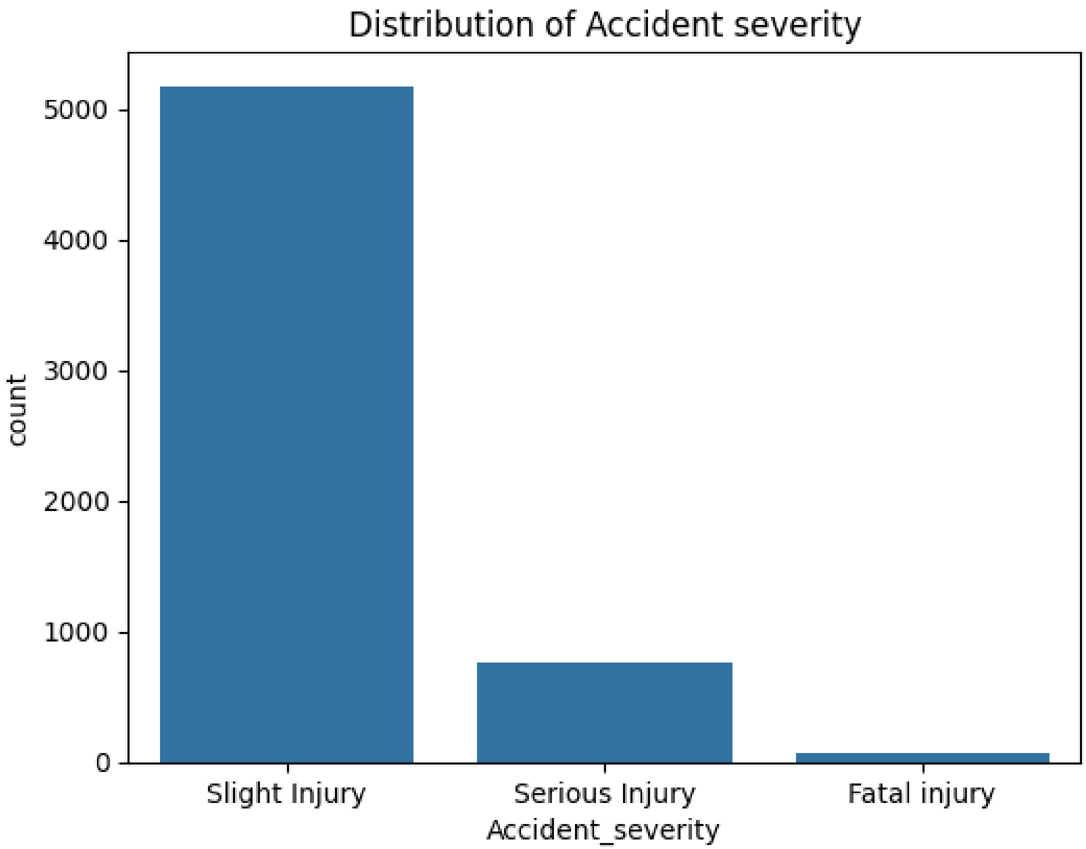
[7]:
```
0
```

[8] :
```
df['Accident_severity'].value_counts()
```

[8] :
```
Accident_severity
Slight Injury     5177
Serious Injury     753
Fatal injury        62
Name: count, dtype: int64
```

[9] :
```
sns.countplot(x = df['Accident_severity'])
plt.title('Distribution of Accident severity')
```

[9] :
```
Text(0.5, 1.0, 'Distribution of Accident severity')
```



[10] :
```
df.isna().sum()
```

[10]:
```
Time                              0
Day_of_week                       0
Age_band_of_driver                0
Sex_of_driver                     1
Educational_level               350
Vehicle_driver_relation         247
Driving_experience              398
Type_of_vehicle                 482
Owner_of_vehicle                231
Service_year_of_vehicle        1974
Defect_of_vehicle              2202
Area_accident_occured           119
Lanes_or_Medians                200
Road_allignment                  68
Types_of_Junction                 1
Road_surface_type                82
Road_surface_conditions           1
Light_conditions                  1
```

```
Weather_conditions                 1
Type_of_collision                 66
Number_of_vehicles_involved        1
Number_of_casualties               1
Vehicle_movement                 123
Casualty_class                     1
Sex_of_casualty                    1
Age_band_of_casualty               1
Casualty_severity                  1
Work_of_casuality               1563
Fitness_of_casuality            1301
Pedestrian_movement                1
Cause_of_accident                  1
Accident_severity                  1
dtype: int64
```

[11]:
```
df.drop(['Service_year_of_vehicle','Defect_of_vehicle','Work_of_casuality',␣
 ↪'Fitness_of_casuality','Time'],
        axis = 1, inplace = True)
df.head()
```

[11]:
```
  Day_of_week Age_band_of_driver Sex_of_driver    Educational_level  \
0      Monday              18-30          Male    Above high school
1      Monday              31-50          Male   Junior high school
2      Monday              18-30          Male   Junior high school
3      Sunday              18-30          Male   Junior high school
4      Sunday              18-30          Male   Junior high school

  Vehicle_driver_relation Driving_experience       Type_of_vehicle  \
0                Employee              1-2yr            Automobile
1                Employee         Above 10yr  Public (> 45 seats)
2                Employee              1-2yr        Lorry (41?100Q)
3                Employee             5-10yr  Public (> 45 seats)
4                Employee              2-5yr                   NaN

  Owner_of_vehicle Area_accident_occured     Lanes_or_Medians  ...  \
0            Owner     Residential areas                  NaN  ...
1            Owner          Office areas   Undivided Two way  ...
2            Owner    Recreational areas                other  ...
3     Governmental          Office areas                other  ...
4            Owner      Industrial areas                other  ...

  Number_of_vehicles_involved Number_of_casualties Vehicle_movement  \
0                         2.0                  2.0   Going straight
1                         2.0                  2.0   Going straight
2                         2.0                  2.0   Going straight
3                         2.0                  2.0   Going straight
```

```
4                          2.0                  2.0    Going straight

    Casualty_class  Sex_of_casualty Age_band_of_casualty Casualty_severity   \
0               na               na                   na                na
1               na               na                   na                na
2  Driver or rider             Male                31-50                 3
3       Pedestrian           Female                18-30                 3
4               na               na                   na                na

   Pedestrian_movement           Cause_of_accident Accident_severity
0     Not a Pedestrian             Moving Backward    Slight Injury
1     Not a Pedestrian                  Overtaking    Slight Injury
2     Not a Pedestrian   Changing lane to the left   Serious Injury
3     Not a Pedestrian  Changing lane to the right    Slight Injury
4     Not a Pedestrian                  Overtaking    Slight Injury

[5 rows x 27 columns]
```

[12] :

```python
categorical=[i for i in df.columns if df[i].dtype=='O']
print('The categorical variables are',categorical)
```

```
The categorical variables are ['Day_of_week', 'Age_band_of_driver',
'Sex_of_driver', 'Educational_level', 'Vehicle_driver_relation',
'Driving_experience', 'Type_of_vehicle', 'Owner_of_vehicle',
'Area_accident_occured', 'Lanes_or_Medians', 'Road_allignment',
'Types_of_Junction', 'Road_surface_type', 'Road_surface_conditions',
'Light_conditions', 'Weather_conditions', 'Type_of_collision',
'Vehicle_movement', 'Casualty_class', 'Sex_of_casualty', 'Age_band_of_casualty',
'Casualty_severity', 'Pedestrian_movement', 'Cause_of_accident',
'Accident_severity']
```

[13] :

```python
for i in categorical:
    df[i].fillna(df[i].mode()[0],inplace=True)
```

[14] :

```python
df.isna().sum()
```

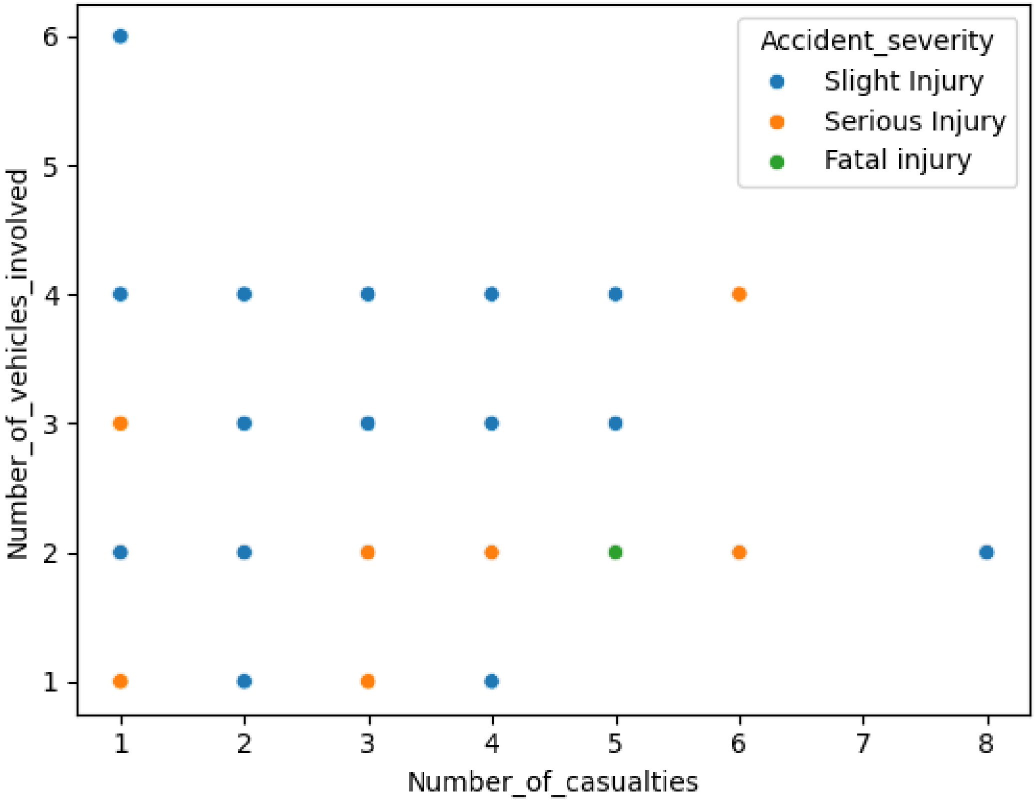
[14]:

```
Day_of_week                0
Age_band_of_driver         0
Sex_of_driver              0
Educational_level          0
Vehicle_driver_relation    0
Driving_experience         0
Type_of_vehicle            0
Owner_of_vehicle           0
Area_accident_occured      0
Lanes_or_Medians           0
Road_allignment            0
```

```
Types_of_Junction              0
Road_surface_type              0
Road_surface_conditions        0
Light_conditions               0
Weather_conditions             0
Type_of_collision              0
Number_of_vehicles_involved    1
Number_of_casualties           1
Vehicle_movement               0
Casualty_class                 0
Sex_of_casualty                0
Age_band_of_casualty           0
Casualty_severity              0
Pedestrian_movement            0
Cause_of_accident              0
Accident_severity              0
dtype: int64
```

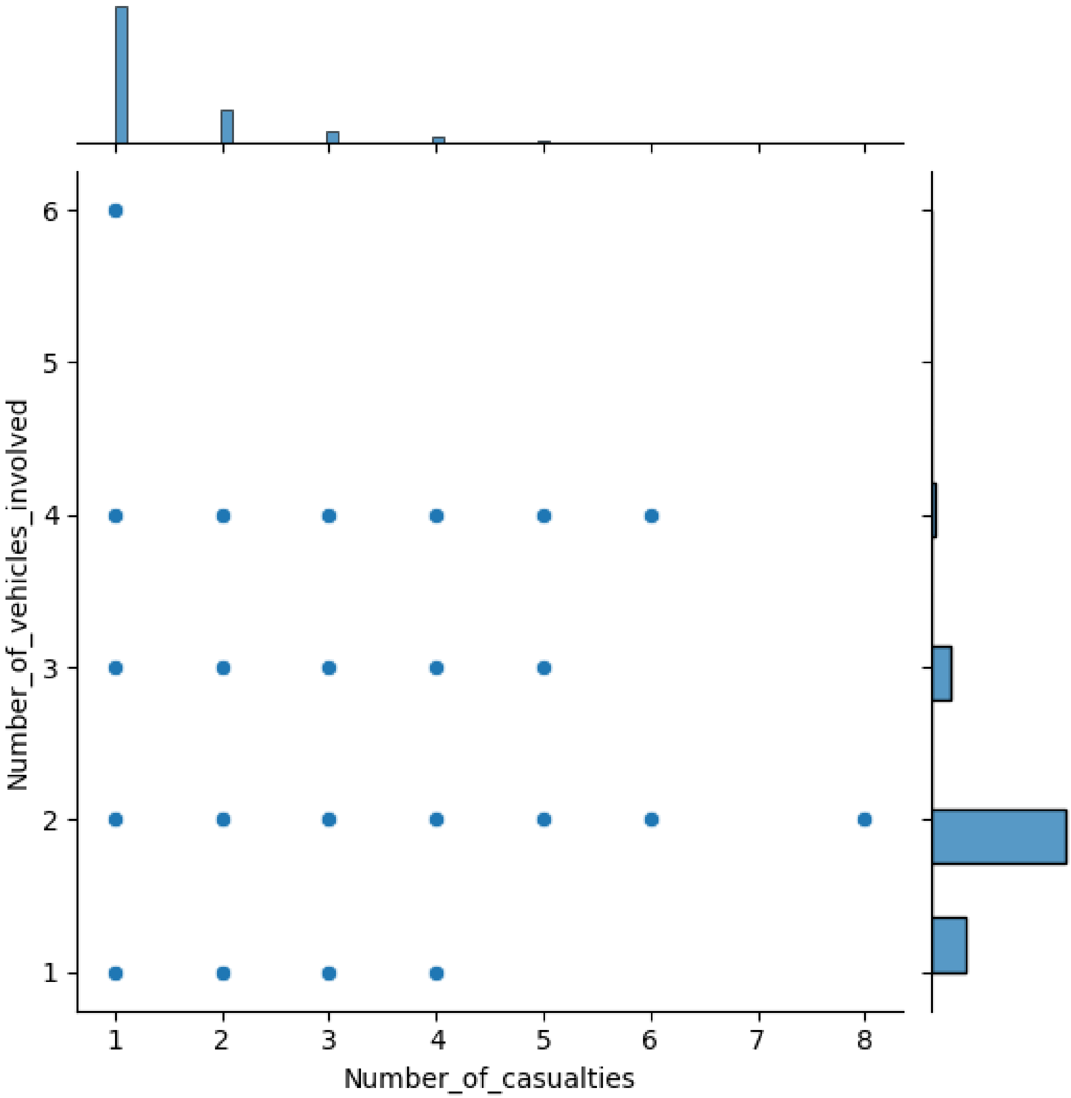
[15]:
```
sns.scatterplot(x=df['Number_of_casualties'],␣
 ↪y=df['Number_of_vehicles_involved'], hue=df['Accident_severity'])
```

[15]:
```
<Axes: xlabel='Number_of_casualties', ylabel='Number_of_vehicles_involved'>
```



[16]:
```
sns.jointplot(x='Number_of_casualties',y='Number_of_vehicles_involved',data=df)
```

[16]:
```
<seaborn.axisgrid.JointGrid at 0x799c62415790>
```

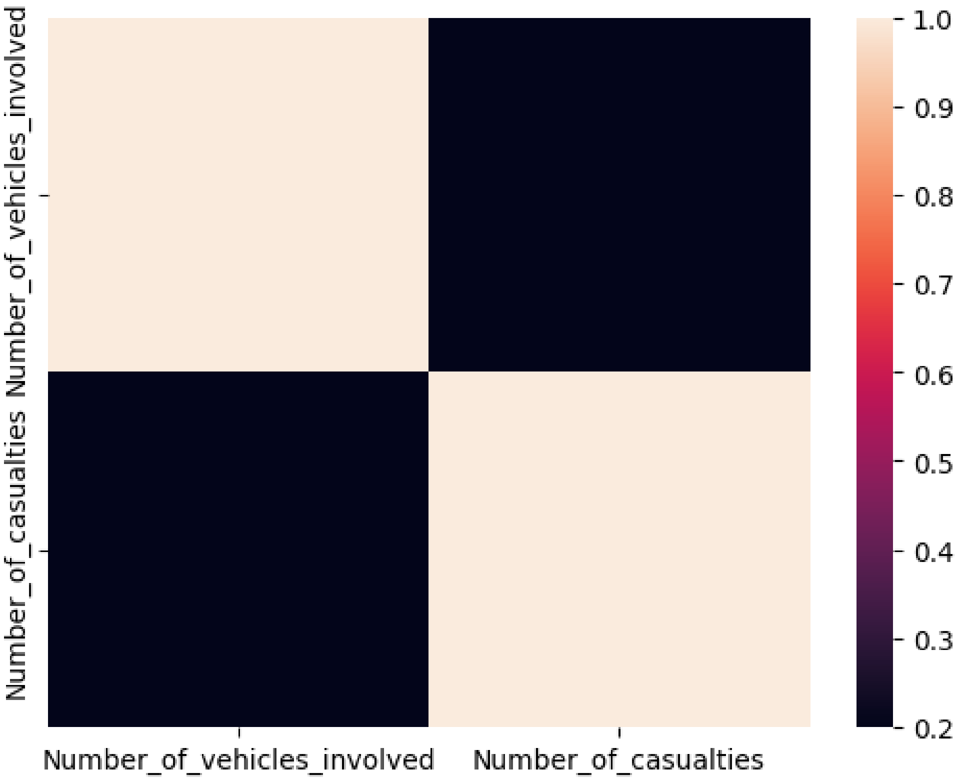


[18]:
```
numerical_df = df.select_dtypes(include=np.number)    # Select numerical columns
correlation_matrix = numerical_df.corr()
print(correlation_matrix)
```

```
                             Number_of_vehicles_involved  Number_of_casualties
Number_of_vehicles_involved                     1.000000              0.199775
Number_of_casualties                            0.199775              1.000000
```

[20]:
```
#plotting the correlation using heatmap
sns.heatmap(df.select_dtypes(include=np.number).corr())
```

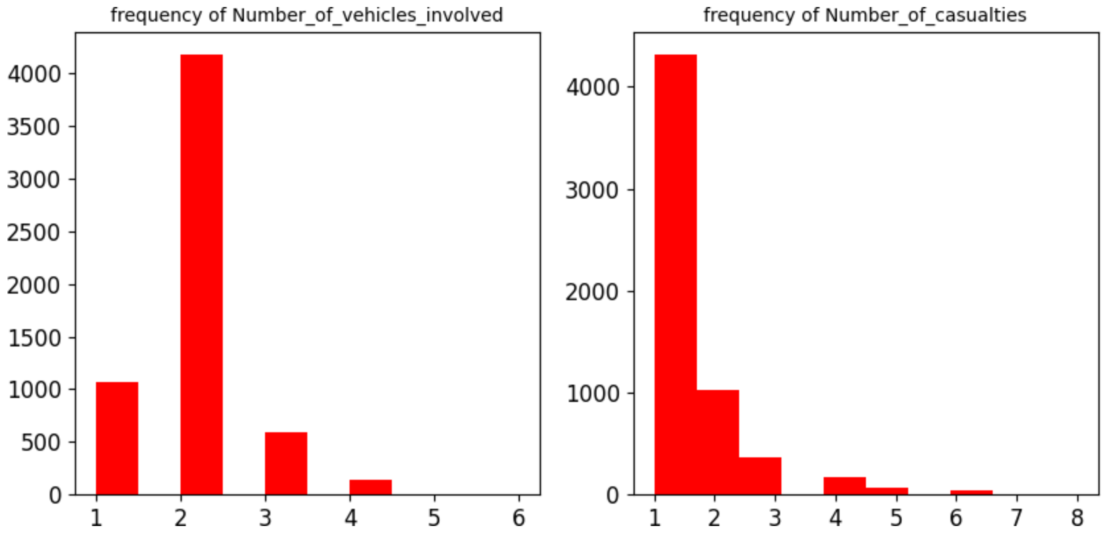
[20] :
```
<Axes: >
```



[21] :
```
numerical=[i for i in df.columns if df[i].dtype!='O']
print('The numerica variables are',numerical)
```

```
The numerica variables are ['Number_of_vehicles_involved',
'Number_of_casualties']
```

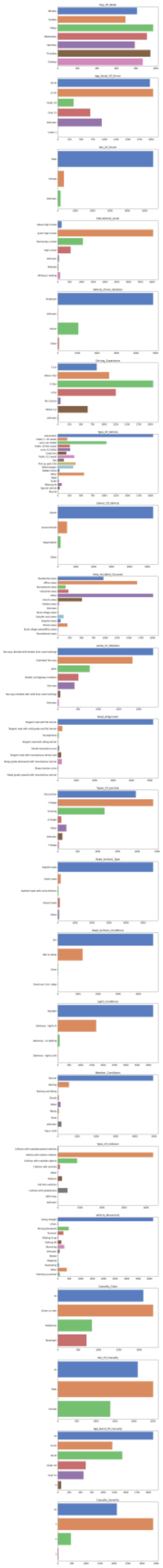
[22] :
```
plt.figure(figsize=(10,10))
plotnumber = 1
for i in numerical:
    if plotnumber <= df.shape[1]:
        ax1 = plt.subplot(2,2,plotnumber)
        plt.hist(df[i],color='red')
        plt.xticks(fontsize=12)
        plt.yticks(fontsize=12)
        plt.title('frequency of '+i, fontsize=10)
    plotnumber +=1
```

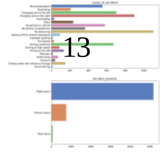


[23] :

```
#count plot for categorical values
plt.figure(figsize=(10,200))
plotnumber = 1

for col in categorical:
    if plotnumber <= df.shape[1] and col!='Pedestrian_movement':
        ax1 = plt.subplot(28,1,plotnumber)
        sns.countplot(data=df, y=col, palette='muted')
        plt.xticks(fontsize=12)
        plt.yticks(fontsize=12)
        plt.title(col.title(), fontsize=14)
        plt.xlabel('')
        plt.ylabel('')
    plotnumber +=1
```


Day_Of_Week
Age_Band_Of_Driver
Sex_Of_Driver
Educational_Level
Vehicle_Driver_Relation
Driving_Experience



[24] :
```
df.dtypes
```

[24] :
```
Day_of_week                    object
Age_band_of_driver             object
Sex_of_driver                  object
Educational_level              object
Vehicle_driver_relation        object
Driving_experience             object
Type_of_vehicle                object
Owner_of_vehicle               object
Area_accident_occured          object
Lanes_or_Medians               object
Road_allignment                object
Types_of_Junction              object
Road_surface_type              object
Road_surface_conditions        object
Light_conditions               object
Weather_conditions             object
Type_of_collision              object
Number_of_vehicles_involved   float64
Number_of_casualties          float64
Vehicle_movement               object
Casualty_class                 object
Sex_of_casualty                object
Age_band_of_casualty           object
Casualty_severity              object
Pedestrian_movement            object
Cause_of_accident              object
Accident_severity              object
dtype: object
```

[25] :
```
#importing label encoing module
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()

#creating a new data frame from performing the chi2 analysis
df1=pd.DataFrame()

#adding all the categorical columns except the output to new data frame
for i in categorical:
    if i!= 'Accident_severity':
        df1[i]=le.fit_transform(df[i])
```

[26] :
```
#confirming the data type
df1.info()
```

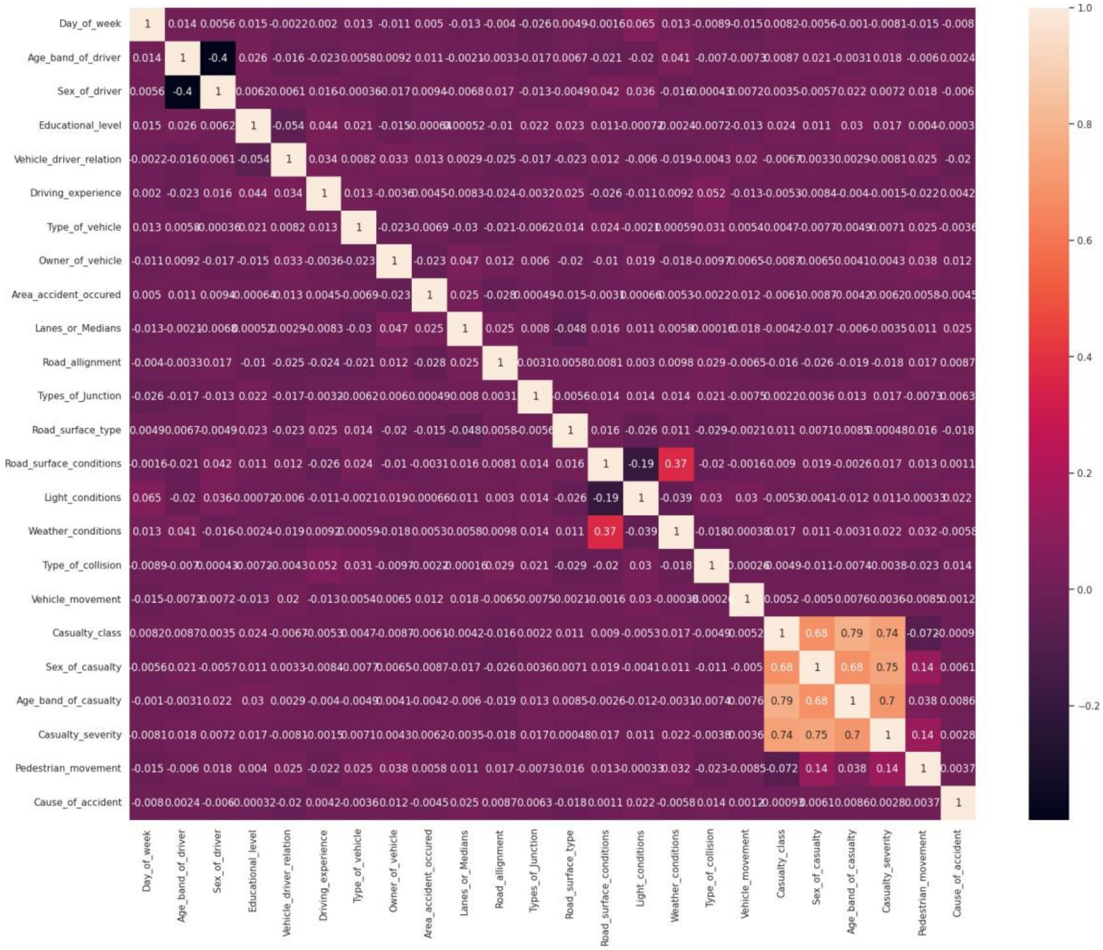
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5993 entries, 0 to 5992
Data columns (total 24 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Day_of_week              5993 non-null   int64
 1   Age_band_of_driver       5993 non-null   int64
 2   Sex_of_driver            5993 non-null   int64
 3   Educational_level        5993 non-null   int64
 4   Vehicle_driver_relation  5993 non-null   int64
 5   Driving_experience       5993 non-null   int64
 6   Type_of_vehicle          5993 non-null   int64
 7   Owner_of_vehicle         5993 non-null   int64
 8   Area_accident_occured    5993 non-null   int64
 9   Lanes_or_Medians         5993 non-null   int64
 10  Road_allignment          5993 non-null   int64
 11  Types_of_Junction        5993 non-null   int64
 12  Road_surface_type        5993 non-null   int64
 13  Road_surface_conditions  5993 non-null   int64
 14  Light_conditions         5993 non-null   int64
 15  Weather_conditions       5993 non-null   int64
 16  Type_of_collision        5993 non-null   int64
 17  Vehicle_movement         5993 non-null   int64
 18  Casualty_class           5993 non-null   int64
 19  Sex_of_casualty          5993 non-null   int64
 20  Age_band_of_casualty     5993 non-null   int64
 21  Casualty_severity        5993 non-null   int64
 22  Pedestrian_movement      5993 non-null   int64
 23  Cause_of_accident        5993 non-null   int64
dtypes: int64(24)
memory usage: 1.1 MB
```

[27]:

```
plt.figure(figsize=(22,17))
sns.set(font_scale=1)
sns.heatmap(df1.corr(), annot=True)
```

[27]:

```
<Axes: >
```



[28] :
```
#label encoded data set
df1.head()
```

[28] :
```
   Day_of_week Age_band_of_driver  Sex_of_driver   Educational_level    \
0            1                  0              1                   0
1            1                  1              1                   4
2            1                  0              1                   4
3            3                  0              1                   4
4            3                  0              1                   4

   Vehicle_driver_relation     Driving_experience  Type_of_vehicle    \
0                        0                      0                 0
1                        0                      3                11
2                        0                      0                 5
3                        0                      2                11
4                        0                      1                 0
```

```
   Owner_of_vehicle  Area_accident_occured  Lanes_or_Medians  ...  \
0                 3                      9                 2  ...
1                 3                      6                 4  ...
2                 3                      1                 6  ...
3                 0                      6                 6  ...
4                 3                      4                 6  ...

   Light_conditions  Weather_conditions  Type_of_collision  Vehicle_movement  \
0                 3                   2                  3                 2
1                 3                   2                  8                 2
2                 3                   2                  2                 2
3                 0                   2                  8                 2
4                 0                   2                  8                 2

   Casualty_class  Sex_of_casualty  Age_band_of_casualty Casualty_severity  \
0               3                2                     5                 3
1               3                2                     5                 3
2               0                1                     1                 2
3               2                0                     0                 2
4               3                2                     5                 3

   Pedestrian_movement  Cause_of_accident
0                    5                  9
1                    5                 16
2                    5                  0
3                    5                  1
4                    5                 16

[5 rows x 24 columns]
```

[29]:
```
#import chi2 test
from sklearn.feature_selection import chi2
f_p_values=chi2(df1,df['Accident_severity'])
```

[30]:
```
#f_p_values will return Fscore and pvalues
f_p_values
```

[30]:
```
(array([ 1.80883262, 9.45043594, 0.01922965, 0.07231767, 13.96771244,
        7.19300683, 0.32360606, 0.73984515, 0.81039617, 3.47759726,
         0.01541708, 7.6266644 , 3.36859125, 4.06800158, 4.31967441,
        1.49649648, 7.48554952, 9.05708919, 0.04410499, 0.51213415,
         7.99699866, 0.15700447, 0.22789597, 2.87309999]),
 array([4.04778081e-01, 8.86878059e-03, 9.90431251e-01, 9.64487088e-01,
        9.26722657e-04, 2.74194294e-02, 8.50608734e-01, 6.90787813e-01,
        6.66844704e-01, 1.75731392e-01, 9.92321095e-01, 2.20744996e-02,
        1.85575100e-01, 1.30811125e-01, 1.15343897e-01, 4.73194750e-01,
        2.36882826e-02, 1.07963778e-02, 9.78188886e-01, 7.74090044e-01,
        1.83431452e-02, 9.24499995e-01, 8.92304370e-01, 2.37746573e-01]))
```

[31] :
```
#for better understanding and ease of access adding them to a new dataframe
f_p_values1=pd.DataFrame({'features':df1.columns, 'Fscore': f_p_values[0],␣
↪'Pvalues':f_p_values[1]})
f_p_values1
```

[31]:
```
                  features     Fscore   Pvalues
0              Day_of_week   1.808833  0.404778
1        Age_band_of_driver   9.450436  0.008869
2            Sex_of_driver   0.019230  0.990431
3         Educational_level   0.072318  0.964487
4   Vehicle_driver_relation  13.967712  0.000927
5        Driving_experience   7.193007  0.027419
6           Type_of_vehicle   0.323606  0.850609
7          Owner_of_vehicle   0.739845  0.690788
8     Area_accident_occured   0.810396  0.666845
9           Lanes_or_Medians   3.477597  0.175731
10           Road_allignment   0.015417  0.992321
11         Types_of_Junction   7.626664  0.022074
12         Road_surface_type   3.368591  0.185575
13   Road_surface_conditions   4.068002  0.130811
14          Light_conditions   4.319674  0.115344
15        Weather_conditions   1.496496  0.473195
16         Type_of_collision   7.485550  0.023688
17          Vehicle_movement   9.057089  0.010796
18            Casualty_class   0.044105  0.978189
19           Sex_of_casualty   0.512134  0.774090
20      Age_band_of_casualty   7.996999  0.018343
21         Casualty_severity   0.157004  0.924500
22       Pedestrian_movement   0.227896  0.892304
23         Cause_of_accident   2.873100  0.237747
```

[32] :
```
#since we want lower Pvalues we are sorting the features
f_p_values1.sort_values(by='Pvalues',ascending=True)
```

[32]:
```
                  features     Fscore   Pvalues
4   Vehicle_driver_relation  13.967712  0.000927
1        Age_band_of_driver   9.450436  0.008869
17          Vehicle_movement   9.057089  0.010796
20      Age_band_of_casualty   7.996999  0.018343
11         Types_of_Junction   7.626664  0.022074
16         Type_of_collision   7.485550  0.023688
5        Driving_experience   7.193007  0.027419
14          Light_conditions   4.319674  0.115344
13   Road_surface_conditions   4.068002  0.130811
9           Lanes_or_Medians   3.477597  0.175731
```

```
12          Road_surface_type   3.368591  0.185575
23          Cause_of_accident   2.873100  0.237747
0                 Day_of_week   1.808833  0.404778
15         Weather_conditions   1.496496  0.473195
8       Area_accident_occured   0.810396  0.666845
7            Owner_of_vehicle   0.739845  0.690788
19             Sex_of_casualty  0.512134  0.774090
6             Type_of_vehicle   0.323606  0.850609
22        Pedestrian_movement   0.227896  0.892304
21          Casualty_severity   0.157004  0.924500
3           Educational_level   0.072318  0.964487
18             Casualty_class   0.044105  0.978189
2               Sex_of_driver   0.019230  0.990431
10            Road_allignment   0.015417  0.992321
```

[33] :

```
#after evaluating we are removing lesser important columns and storing to a new␣
↪data frame
df2=df.drop(['Owner_of_vehicle', 'Type_of_vehicle', 'Road_surface_conditions',␣
↪'Pedestrian_movement',
           ␣
↪'Casualty_severity','Educational_level','Day_of_week','Sex_of_driver','Road_allignment',
            'Sex_of_casualty'],axis=1)
df2.head()
```

[33]:

```
  Age_band_of_driver Vehicle_driver_relation Driving_experience  \
0              18-30                Employee              1-2yr
1              31-50                Employee         Above 10yr
2              18-30                Employee              1-2yr
3              18-30                Employee             5-10yr
4              18-30                Employee              2-5yr

  Area_accident_occured                                  Lanes_or_Medians  \
0     Residential areas  Two-way (divided with broken lines road marking)
1          Office areas                                   Undivided Two way
2     Recreational areas                                              other
3          Office areas                                              other
4      Industrial areas                                              other

  Types_of_Junction Road_surface_type        Light_conditions  \
0       No junction    Asphalt roads                 Daylight
1       No junction    Asphalt roads                 Daylight
2       No junction    Asphalt roads                 Daylight
3           Y Shape      Earth roads  Darkness - lights lit
4           Y Shape    Asphalt roads  Darkness - lights lit

  Weather_conditions                          Type_of_collision  \
0             Normal  Collision with roadside-parked vehicles
```

```
1           Normal         Vehicle with vehicle collision
2           Normal       Collision with roadside objects
3           Normal         Vehicle with vehicle collision
4           Normal         Vehicle with vehicle collision

   Number_of_vehicles_involved   Number_of_casualties Vehicle_movement  \
0                          2.0                    2.0    Going straight
1                          2.0                    2.0    Going straight
2                          2.0                    2.0    Going straight
3                          2.0                    2.0    Going straight
4                          2.0                    2.0    Going straight

   Casualty_class  Age_band_of_casualty            Cause_of_accident  \
0              na                    na              Moving Backward
1              na                    na                   Overtaking
2  Driver or rider                31-50    Changing lane to the left
3       Pedestrian                18-30   Changing lane to the right
4              na                    na                   Overtaking

  Accident_severity
0     Slight Injury
1     Slight Injury
2    Serious Injury
3     Slight Injury
4     Slight Injury
```

[34] :

```
df2.shape
```

[34]:

```
(5993, 17)
```

[35] :

```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5993 entries, 0 to 5992
Data columns (total 17 columns):
 #   Column                       Non-Null Count  Dtype
---  -------                      --------------  ------
 0   Age_band_of_driver           5993 non-null   object
 1   Vehicle_driver_relation      5993 non-null   object
 2   Driving_experience           5993 non-null   object
 3   Area_accident_occured        5993 non-null   object
 4   Lanes_or_Medians             5993 non-null   object
 5   Types_of_Junction            5993 non-null   object
 6   Road_surface_type            5993 non-null   object
 7   Light_conditions             5993 non-null   object
 8   Weather_conditions           5993 non-null   object
 9   Type_of_collision            5993 non-null   object
 10  Number_of_vehicles_involved  5992 non-null   float64
 11  Number_of_casualties         5992 non-null   float64
 12  Vehicle_movement             5993 non-null   object
 13  Casualty_class               5993 non-null   object
 14  Age_band_of_casualty         5993 non-null   object
 15  Cause_of_accident            5993 non-null   object
 16  Accident_severity            5993 non-null   object
dtypes: float64(2), object(15)
memory usage: 796.1+ KB
```

```
[36]: #to check distinct values in each categorical columns we are storing them to a new variable
categorical_new=[i for i in df2.columns if df2[i].dtype=='O']
print(categorical_new)
```

```
['Age_band_of_driver', 'Vehicle_driver_relation', 'Driving_experience',
'Area_accident_occured', 'Lanes_or_Medians', 'Types_of_Junction',
'Road_surface_type', 'Light_conditions', 'Weather_conditions',
'Type_of_collision', 'Vehicle_movement', 'Casualty_class',
'Age_band_of_casualty', 'Cause_of_accident', 'Accident_severity']
```

```
[37]: for i in categorical_new:
    print(df2[i].value_counts())
```

```
Age_band_of_driver
31-50       2042
18-30       1974
Unknown      942
Over 51      695
Under 18     339
Under 1        1
Name: count, dtype: int64
Vehicle_driver_relation
Employee    4877
Owner       1042
Other         60
Unknown       14
Name: count, dtype: int64
Driving_experience
5-10yr        2063
2-5yr         1254
Above 10yr    1111
1-2yr          841
Below 1yr      646
No Licence      60
unknown         18
Name: count, dtype: int64
Area_accident_occured
```

```
Other                                2012
Office  areas                        1674
Residential  areas                    964
 Church areas                         510
 Industrial  areas                    223
School areas                          203
   Recreational  areas                164
 Outside rural areas                  119
 Hospital  areas                       59
   Market areas                        29
Rural village areas                    18
Unknown                                 9
Rural village areasOffice areas         8
Recreational  areas                     1
Name: count, dtype: int64
Lanes_or_Medians
Two-way (divided with broken lines   road marking)     2356
Undivided Two way                                      1844
other                                                   788
Double carriageway (median)                             507
One way                                                 411
Two-way (divided with solid lines road marking)          62
Unknown                                                  25
Name: count, dtype: int64
Types_of_Junction
Y Shape          2400
No junction      1966
Crossing         1178
Other             215
Unknown           115
O Shape            89
T Shape            30
Name: count, dtype: int64
Road_surface_type
Asphalt roads                          5587
Earth  roads                            169
Gravel  roads                           117
Other                                    80
Asphalt roads with some distress         40
Name: count, dtype: int64
Light_conditions
Daylight                       4206
Darkness - lights lit          1693
Darkness - no lighting           83
Darkness - lights unlit          11
Name: count, dtype: int64
Weather_conditions
Normal                 4949
```

```
Raining                578
Unknown                190
Other                  139
Cloudy                  48
Windy                   43
Snow                    21
Raining and Windy       20
Fog or mist              5
Name: count, dtype: int64
Type_of_collision
Vehicle with vehicle collision                 4319
Collision with roadside objects                 871
Collision with pedestrians                      440
Rollover                                        189
Collision with animals                           95
Collision with roadside-parked vehicles          30
Fall from vehicles                               17
Unknown                                          14
Other                                            13
With Train                                        5
Name: count, dtype:  int64
Vehicle_movement
Going straight            4156
Moving Backward            475
Other                      389
Reversing                  281
Turnover                   243
Getting off                152
Entering a junction         86
Unknown                     81
Overtaking                  44
Stopping                    29
Waiting to go               25
U-Turn                      22
Parked                      10
Name: count, dtype:  int64
Casualty_class
Driver or rider      2345
na                   2105
Pedestrian            838
Passenger             705
Name: count, dtype: int64
Age_band_of_casualty
na             2106
18-30          1423
31-50          1204
Under 18        618
Over 51         567
```

```
5                 75
Name: count, dtype: int64
Cause_of_accident
No distancing                           1105
Changing lane to the right               898
Changing lane to the left                704
Driving carelessly                       672
No priority to vehicle                   579
Moving Backward                          553
No priority to pedestrian                365
Other                                    235
Overtaking                               208
Driving under the influence of drugs     151
Driving to the left                      135
Getting off the vehicle improperly        92
Driving at high speed                     87
Overturning                               71
Turnover                                  40
Overloading                               30
Overspeed                                 28
Drunk driving                             14
Unknown                                   13
Improper parking                          13
Name: count, dtype: int64
Accident_severity
Slight Injury     5178
Serious Injury     753
Fatal injury        62
Name: count, dtype: int64
```

[39]:

```
#get_dummies
dummy=pd.get_dummies(df2[['Age_band_of_driver',  'Vehicle_driver_relation',␣
 ↪'Driving_experience',
                            'Area_accident_occured', 'Lanes_or_Medians',␣
 ↪'Types_of_Junction', 'Road_surface_type',
                            'Light_conditions',  'Weather_conditions',␣
 ↪'Type_of_collision', 'Vehicle_movement',
                            'Casualty_class', 'Age_band_of_casualty',␣
 ↪'Cause_of_accident']],drop_first=True)
dummy.head()
```

[39]:

```
   Age_band_of_driver_31-50  Age_band_of_driver_Over 51  \
0                     False                       False
1                      True                       False
2                     False                       False
3                     False                       False
4                     False                       False
```

```
   Age_band_of_driver_Under 1   Age_band_of_driver_Under 18  \
0                       False                         False
1                       False                         False
2                       False                         False
3                       False                         False
4                       False                         False

   Age_band_of_driver_Unknown Vehicle_driver_relation_Other     \
0                       False                         False
1                       False                         False
2                       False                         False
3                       False                         False
4                       False                         False

   Vehicle_driver_relation_Owner    Vehicle_driver_relation_Unknown    \
0                          False                               False
1                          False                               False
2                          False                               False
3                          False                               False
4                          False                               False

   Driving_experience_2-5yr    Driving_experience_5-10yr   ...  \
0                     False                        False   ...
1                     False                        False   ...
2                     False                        False   ...
3                     False                         True   ...
4                      True                        False   ...

   Cause_of_accident_No distancing   \
0                            False
1                            False
2                            False
3                            False
4                            False

   Cause_of_accident_No  priority  to  pedestrian   \
0                                           False
1                                           False
2                                           False
3                                           False
4                                           False

   Cause_of_accident_No  priority  to  vehicle   Cause_of_accident_Other   \
0                                       False                     False
1                                       False                     False
2                                       False                     False
```

```
3                                      False                      False
4                                      False                      False

   Cause_of_accident_Overloading  Cause_of_accident_Overspeed  \
0                          False                        False
1                          False                        False
2                          False                        False
3                          False                        False
4                          False                        False

   Cause_of_accident_Overtaking  Cause_of_accident_Overturning  \
0                         False                          False
1                          True                          False
2                         False                          False
3                         False                          False
4                          True                          False

   Cause_of_accident_Turnover  Cause_of_accident_Unknown
0                       False                      False
1                       False                      False
2                       False                      False
3                       False                      False
4                       False                      False

[5 rows x 102 columns]
```

[40]:

```
#concatinate dummy and old data frame
df3=pd.concat([df2,dummy],axis=1)
df3.head()
```

[40]:

```
  Age_band_of_driver Vehicle_driver_relation Driving_experience  \
0              18-30                Employee              1-2yr
1              31-50                Employee         Above 10yr
2              18-30                Employee              1-2yr
3              18-30                Employee             5-10yr
4              18-30                Employee              2-5yr

  Area_accident_occured                                   Lanes_or_Medians  \
0     Residential areas  Two-way (divided with broken lines road marking)
1          Office areas                                   Undivided Two way
2    Recreational areas                                               other
3          Office areas                                               other
4      Industrial areas                                               other

  Types_of_Junction Road_surface_type       Light_conditions  \
0       No junction     Asphalt roads               Daylight
1       No junction     Asphalt roads               Daylight
```

```
2          No junction       Asphalt roads                  Daylight
3              Y Shape         Earth roads   Darkness - lights lit
4              Y Shape       Asphalt roads   Darkness - lights lit

  Weather_conditions                              Type_of_collision   ...  \
0             Normal  Collision with roadside-parked vehicles   ...
1             Normal             Vehicle with vehicle collision   ...
2             Normal           Collision with roadside objects   ...
3             Normal             Vehicle with vehicle collision   ...
4             Normal             Vehicle with vehicle collision   ...

   Cause_of_accident_No distancing   \
0                             False
1                             False
2                             False
3                             False
4                             False

   Cause_of_accident_No priority to pedestrian   \
0                                         False
1                                         False
2                                         False
3                                         False
4                                         False

  Cause_of_accident_No priority to vehicle Cause_of_accident_Other  \
0                                    False                   False
1                                    False                   False
2                                    False                   False
3                                    False                   False
4                                    False                   False

  Cause_of_accident_Overloading Cause_of_accident_Overspeed  \
0                         False                       False
1                         False                       False
2                         False                       False
3                         False                       False
4                         False                       False

  Cause_of_accident_Overtaking   Cause_of_accident_Overturning   \
0                        False                          False
1                         True                          False
2                        False                          False
3                        False                          False
4                         True                          False

   Cause_of_accident_Turnover   Cause_of_accident_Unknown
```

```
0                            False                          False
1                            False                          False
2                            False                          False
3                            False                          False
4                            False                          False

[5 rows x 119 columns]
```

[41] :

```
#dropping dummied columns
df3.drop(['Age_band_of_driver', 'Vehicle_driver_relation',␣
↪'Driving_experience', 'Area_accident_occured', 'Lanes_or_Medians',
          'Types_of_Junction', 'Road_surface_type', 'Light_conditions',␣
↪'Weather_conditions', 'Type_of_collision',
          'Vehicle_movement','Casualty_class', 'Age_band_of_casualty',␣
↪'Cause_of_accident'],axis=1,inplace=True)
df3.head()
```

[41]:

```
   Number_of_vehicles_involved  Number_of_casualties  Accident_severity  \
0                          2.0                   2.0     Slight Injury
1                          2.0                   2.0     Slight Injury
2                          2.0                   2.0    Serious Injury
3                          2.0                   2.0     Slight Injury
4                          2.0                   2.0     Slight Injury

   Age_band_of_driver_31-50  Age_band_of_driver_Over 51  \
0                     False                       False
1                      True                       False
2                     False                       False
3                     False                       False
4                     False                       False

   Age_band_of_driver_Under 1  Age_band_of_driver_Under 18  \
0                       False                        False
1                       False                        False
2                       False                        False
3                       False                        False
4                       False                        False

   Age_band_of_driver_Unknown Vehicle_driver_relation_Other  \
0                       False                         False
1                       False                         False
2                       False                         False
3                       False                         False
4                       False                         False

   Vehicle_driver_relation_Owner  … Cause_of_accident_No distancing  \
0                          False  …                           False
```

```
1                          False  ...                          False
2                          False  ...                          False
3                          False  ...                          False
4                          False  ...                          False

   Cause_of_accident_No priority to pedestrian  \
0                                        False
1                                        False
2                                        False
3                                        False
4                                        False

   Cause_of_accident_No priority to vehicle  Cause_of_accident_Other  \
0                                     False                    False
1                                     False                    False
2                                     False                    False
3                                     False                    False
4                                     False                    False

   Cause_of_accident_Overloading  Cause_of_accident_Overspeed  \
0                          False                        False
1                          False                        False
2                          False                        False
3                          False                        False
4                          False                        False

   Cause_of_accident_Overtaking  Cause_of_accident_Overturning  \
0                         False                          False
1                          True                          False
2                         False                          False
3                         False                          False
4                          True                          False

   Cause_of_accident_Turnover  Cause_of_accident_Unknown
0                       False                      False
1                       False                      False
2                       False                      False
3                       False                      False
4                       False                      False

[5 rows x 105 columns]
```

[42] :

```
x=df3.drop(['Accident_severity'],axis=1)
x.shape
```

[42]:

```
(5993, 104)
```

[43] :
```
x.head()
```

[43]:
```
   Number_of_vehicles_involved  Number_of_casualties  \
0                          2.0                   2.0
1                          2.0                   2.0
2                          2.0                   2.0
3                          2.0                   2.0
4                          2.0                   2.0

   Age_band_of_driver_31-50  Age_band_of_driver_Over 51  \
0                     False                       False
1                      True                       False
2                     False                       False
3                     False                       False
4                     False                       False

   Age_band_of_driver_Under 1  Age_band_of_driver_Under 18  \
0                       False                        False
1                       False                        False
2                       False                        False
3                       False                        False
4                       False                        False

   Age_band_of_driver_Unknown Vehicle_driver_relation_Other  \
0                       False                         False
1                       False                         False
2                       False                         False
3                       False                         False
4                       False                         False

   Vehicle_driver_relation_Owner  Vehicle_driver_relation_Unknown  ...  \
0                          False                            False  ...
1                          False                            False  ...
2                          False                            False  ...
3                          False                            False  ...
4                          False                            False  ...

   Cause_of_accident_No distancing  \
0                            False
1                            False
2                            False
3                            False
4                            False

   Cause_of_accident_No priority to pedestrian  \
0                                        False
1                                        False
```

```
2                                    False
3                                    False
4                                    False

   Cause_of_accident_No priority to vehicle  Cause_of_accident_Other  \
0                                     False                    False
1                                     False                    False
2                                     False                    False
3                                     False                    False
4                                     False                    False

   Cause_of_accident_Overloading  Cause_of_accident_Overspeed  \
0                          False                        False
1                          False                        False
2                          False                        False
3                          False                        False
4                          False                        False

   Cause_of_accident_Overtaking  Cause_of_accident_Overturning  \
0                         False                          False
1                          True                          False
2                         False                          False
3                         False                          False
4                          True                          False

   Cause_of_accident_Turnover  Cause_of_accident_Unknown
0                       False                      False
1                       False                      False
2                       False                      False
3                       False                      False
4                       False                      False

[5 rows x 104 columns]
```

[44]:

```
y=df3.iloc[:,2]
y.head()
```

[44]:

```
0     Slight Injury
1     Slight Injury
2    Serious Injury
3     Slight Injury
4     Slight Injury
Name: Accident_severity, dtype: object
```

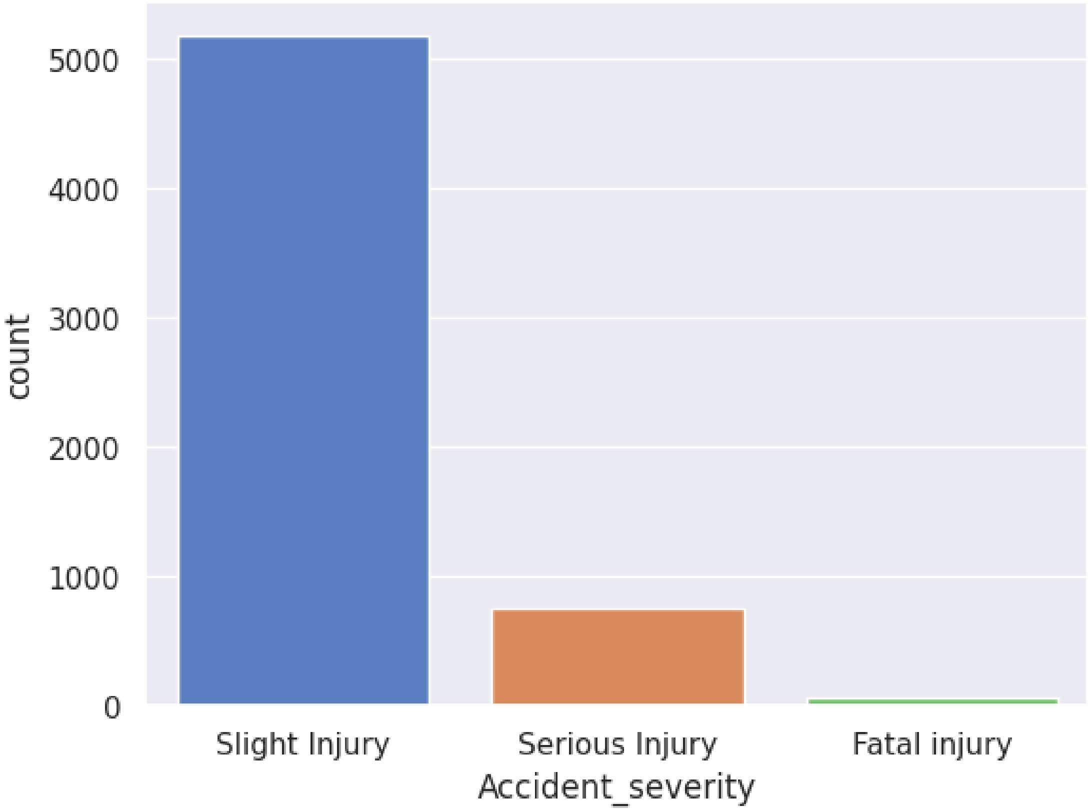
[45]:

```
#checking the count of each item in the output column
y.value_counts()
```

[45]:
```
Accident_severity
Slight Injury      5178
Serious Injury      753
Fatal injury         62
Name: count, dtype: int64
```

[46]:
```python
#plotting count plot using seaborn
sns.countplot(x = y, palette='muted')
```

[46]:
```
<Axes: xlabel='Accident_severity', ylabel='count'>
```



[49]:
```python
# Impute missing values using SimpleImputer
from sklearn.impute import SimpleImputer

# Create an imputer object with your desired strategy (e.g., mean, median,␣
↪most_frequent)
imputer = SimpleImputer(strategy='most_frequent')    # Replace with your␣
↪preferred strategy

# Fit the imputer on your data and transform it
x_imputed = imputer.fit_transform(x)
```

```
# Now, apply SMOTE on the imputed data
xo, yo = oversample.fit_resample(x_imputed, y)
```

[50]:

```
#checking the oversampling output
y1=pd.DataFrame(yo)
y1.value_counts()
```

[50]:

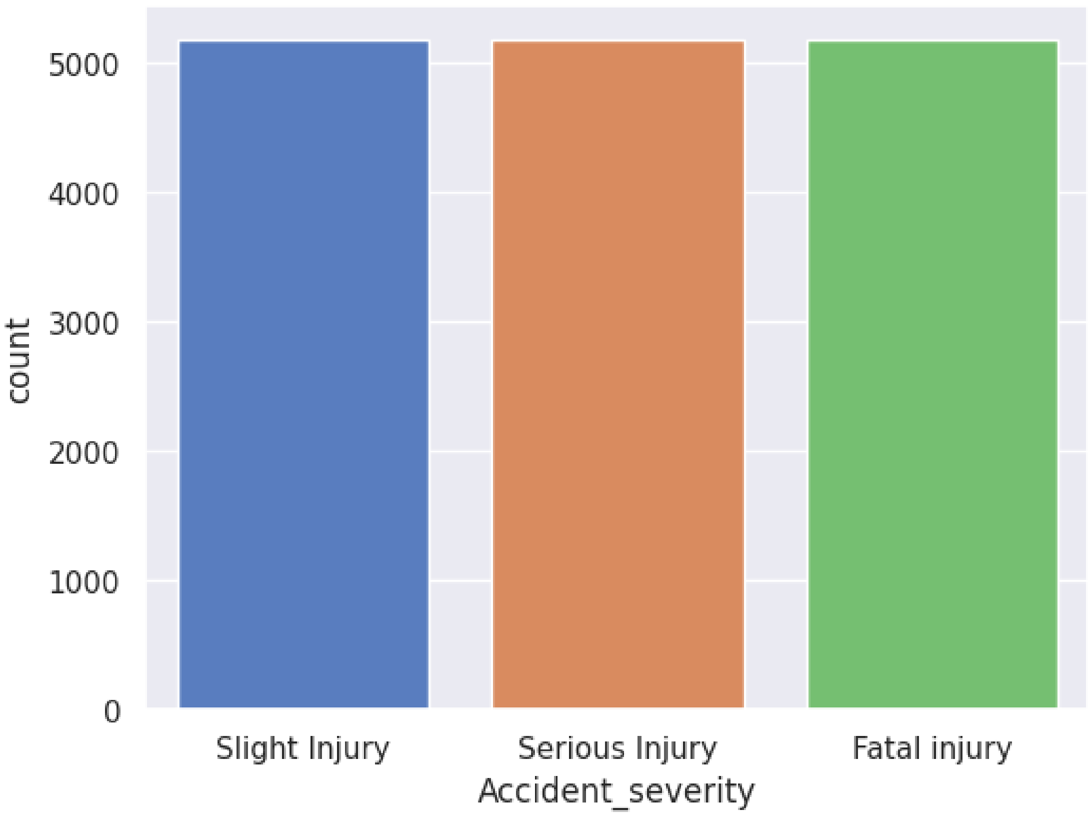
```
Accident_severity
Fatal  injury          5178
Serious  Injury        5178
Slight Injury          5178
Name: count, dtype: int64
```

[51]:

```
sns.countplot(x = yo, palette='muted')
```

[51] :

```
<Axes:  xlabel='Accident_severity',  ylabel='count'>
```



[52] :

```
#converting data to training data and testing data
from sklearn.model_selection import train_test_split
#splitting 70% of the data to training data and 30% of data to testing data
```

```
x_train,x_test,y_train,y_test=train_test_split(xo,yo,test_size=0.
↪30,random_state=42)
```

[53]:

```
print(x_train.shape,x_test.shape,y_train.shape,y_test.shape)
```

```
(10873, 104) (4661, 104) (10873,) (4661,)
```

[54]:

```
#KNN model alg
from sklearn.neighbors import KNeighborsClassifier
model_KNN=KNeighborsClassifier(n_neighbors=5)
model_KNN.fit(x_train,y_train)
```

[54]:

```
KNeighborsClassifier()
```

[55]:

```
y_pred=model_KNN.predict(x_test)
```

[56]:

```
y_pred
```

[56]:

```
array(['Serious Injury', 'Serious Injury', 'Serious Injury', …,
       'Serious Injury', 'Fatal injury', 'Fatal injury'], dtype=object)
```

[57]:

```
from sklearn.metrics import␣
↪classification_report,confusion_matrix,accuracy_score,ConfusionMatrixDisplay
```

[58]:

```
report_KNN=classification_report(y_test,y_pred)
print(report_KNN)
```

```
                precision    recall  f1-score   support

  Fatal injury       0.88      1.00      0.94      1548
Serious Injury       0.67      0.99      0.80      1551
 Slight Injury       0.99      0.39      0.56      1562

      accuracy                           0.79      4661
     macro avg       0.85      0.79      0.77      4661
  weighted avg       0.85      0.79      0.77      4661
```

[59]:

```
accuracy_KNN=accuracy_score(y_test,y_pred)
print(accuracy_KNN)
```

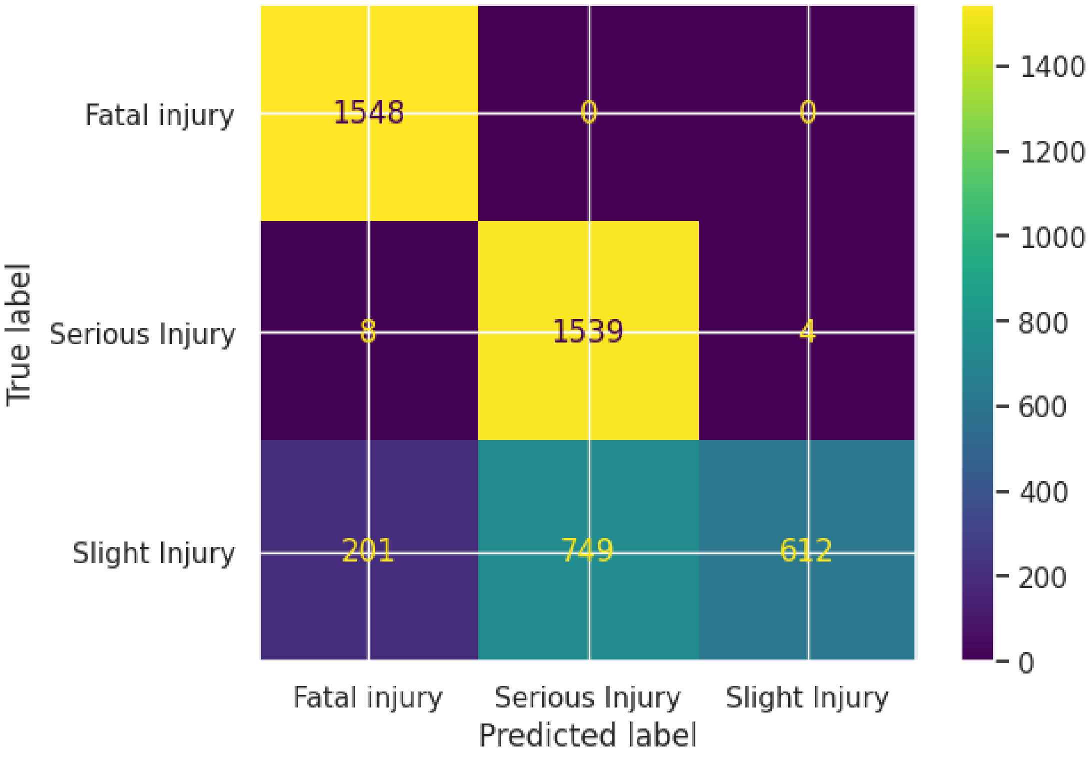
```
0.7936065222055353
```

[60]:

```
matrix_KNN=confusion_matrix(y_test,y_pred)
print(matrix_KNN,'\n')
print(ConfusionMatrixDisplay.from_predictions(y_test,y_pred))
```

```
[[1548    0    0]
 [   8 1539    4]
 [ 201  749  612]]

<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay object at
0x799c558e1f50>
```


True label
Fatal injury
Serious Injury
Slight Injury
Fatal injury
Serious Injury
Slight Injury
Predicted label