ROAD ACCIDENTS DATA ANALYSIS

By Nikhil Kaundal

Problem Statement

The **Problem Statement** for this project is to find out the which factors are associated with higher number of accidents and which groups are at a higher end in getting into those accidents.

Objectives of EDA : The objectives of performing EDA on this dataset is

- To get a understanding of the data.
- To identify pattern and trends in the data.
- To use data visualization techniques like histograms, box plots, and heatmaps to represent data visually.
- To develop hypothesis which can be further tested for a more targeted data analysis.

Dataset Overview

This dataset contains information about road accidents that occurred in United Kingdom in the year 2022. A total of number of 61,352 accidents are listed in this dataset.

The dataset encompasses various attributes related to accident status, vehicle and casualty references, demographics, and severity of casualties. It includes essential factors such as pedestrian details, casualty types, road maintenance worker involvement, and the Index of Multiple Deprivation (IMD) decile for casualties' home areas.

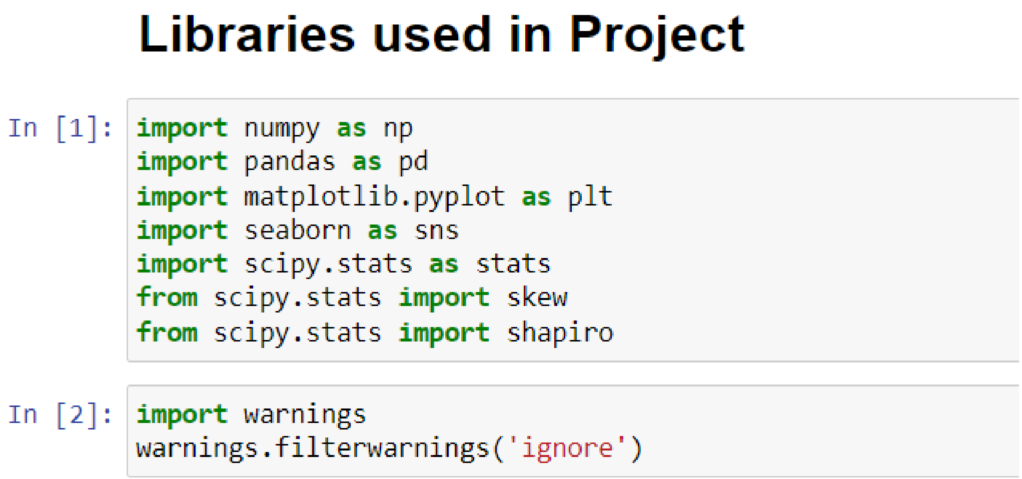
The dataset included 61,352 samples and 20 features.

Data source: This dataset was collected from Kaggle website https://www.kaggle.com/datasets/juhibhojani/road-accidents-data-2022

Libraries Used

The libraries that I have used in this project are:

NumPy, Pandas, Matplotlib, Seaborn, SciPy



Libraries used in Project

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        import scipy.stats as stats
        from scipy.stats import skew
        from scipy.stats import shapiro

In [2]: import warnings
        warnings.filterwarnings('ignore')
```

Approach to solve the problem

1. First of all import a dataset from Kaggle or UCI Machine Learning Repository.
2. Then implement summary statistics see what are the mean standard deviation, min max of columns of the dataset.
3. Through summary we can see that some columns have negative value for some records.
4. After removing those invalid values univariate analysis can be implemented through bar graph and histogram.
5. After getting insight from univariate , bivariate analysis and multivariate analysis can also be implemented which will give us detail how more than one variables are related to each other.
6. Then using histplot() visualize which distribution is the dataset following.
7. At last develop some hypothesis and test whether the dataset is following those hypothesis or not.

Data Cleaning

Before starting the analysis, first we need to clean and filter the data.

- ***Check null values***

  No null values were found.

- ***Drop duplicates if any***

  No duplicates were found.

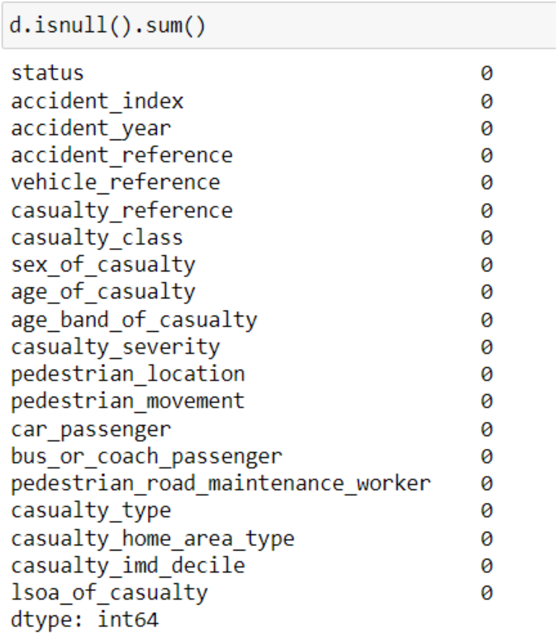
- ***Drop unwanted columns***

  status, accident_index, accident_year were dropped.

- ***Remove invalid values***

  Invalid values were removed from sex_of_casualty, age_of_casualty,, casualty_type, casualty_home_area_type,casualty_imd_decile and other columns

After removing invalid values the size of data points were reduced to 54,281 from 61,352.



```
d.isnull().sum()
```

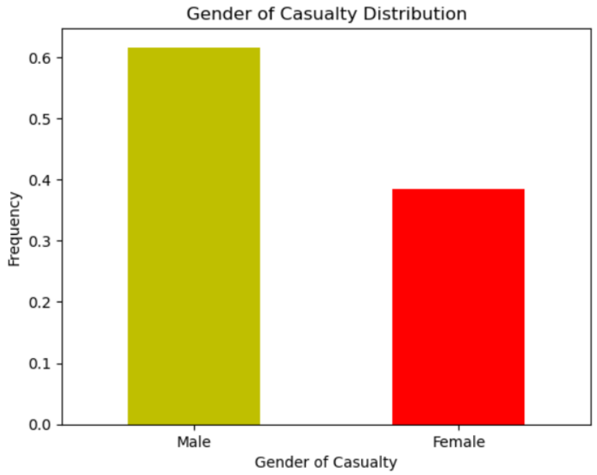
```
status                                 0
accident_index                         0
accident_year                          0
accident_reference                     0
vehicle_reference                      0
casualty_reference                     0
casualty_class                         0
sex_of_casualty                        0
age_of_casualty                        0
age_band_of_casualty                   0
casualty_severity                      0
pedestrian_location                    0
pedestrian_movement                    0
car_passenger                          0
bus_or_coach_passenger                 0
pedestrian_road_maintenance_worker     0
casualty_type                          0
casualty_home_area_type                0
casualty_imd_decile                    0
lsoa_of_casualty                       0
dtype: int64
```

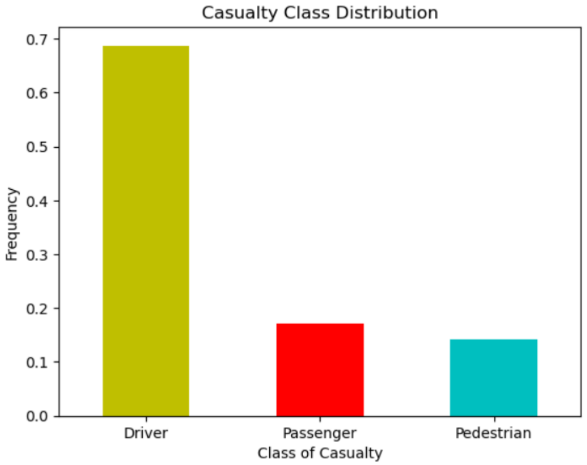
Univariate Analysis

Univariate Analysis was performed using bargraphs and histograms.

Columns on which univariate analysis was performed.

Column	*Insight*
sex_of_casualty	Male
age_band_of_casualty	*Aged(26-35)*
casualty_class	*Driver*
casualty_severity	*Slight*
casualty_home_area_type	*Urban*
casualty_imd_decile	*2nd IMD decile*



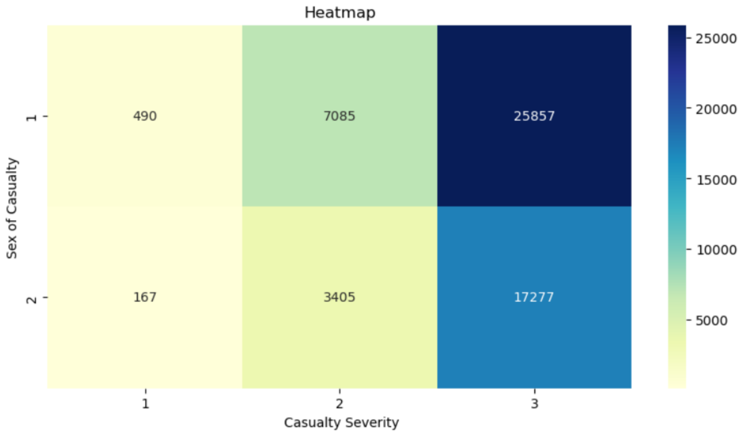


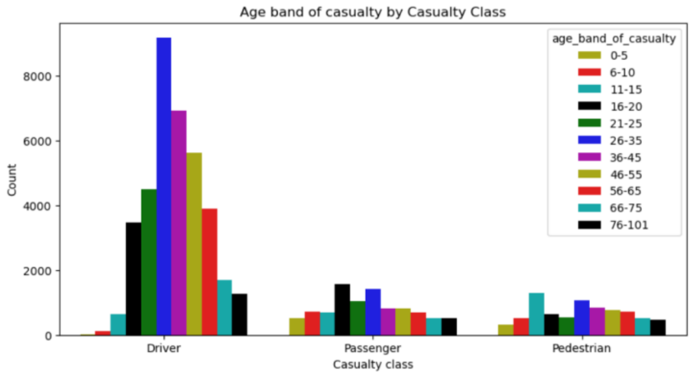
Bivariate Analysis

Bivariate Analysis was performed using countplot and heatmap

Columns on which bivariate analysis was performed.

- ***casualty_severity & casualty_class***
  - Driver has highest no of casualty in any severity.
- ***casualty_severity & sex_of_casualty***
  - Male has highest no of casualty in any severity.
- ***casualty_severity & age_band_of_casualty***
  - Age group(26-35) has highest no of casualty in any severity
- ***casualty_severity & casualty_home_area_type***
  - *Urban area* has highest no of casualty in any severity
- ***casualty_severity & casualty_imd_decile***
  - No of casualties decrease as imd decile improves for all severities.
- ***casualty_class & age_band_of_casualty***
  - For driver the graph increases from 0-5 to 26-35 age group and decreases for higher age groups
- ***casualty_class & sex_of_casualty***
  - Most driver & pedestrian caught in an accident are male



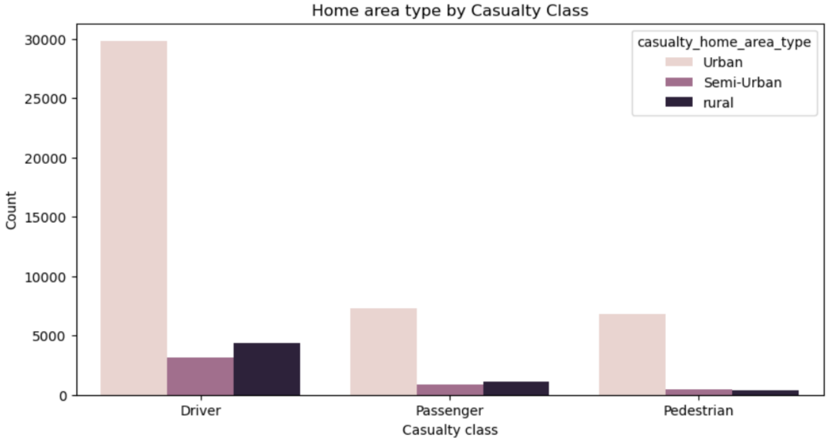


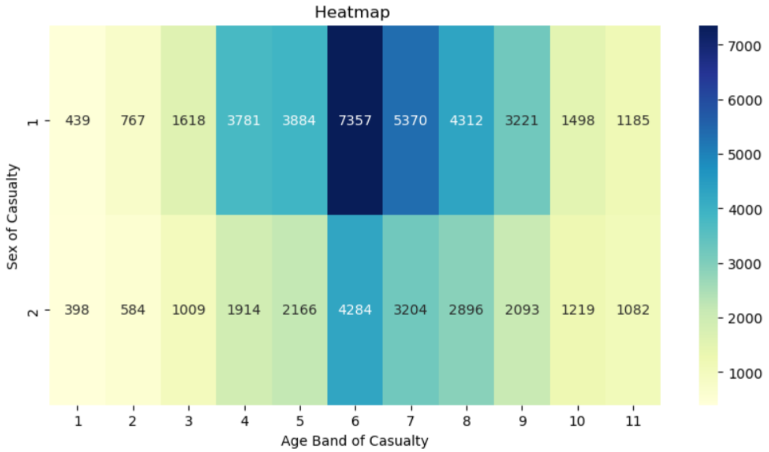
Bivariate Analysis

Univariate Analysis was performed using countplot and heatmap

Columns on which bivariate analysis was performed.

- ***casualty_class & casualty_home_area_type***
  - Irrespective of class most casualties are from urban area
- ***sex_of_casualty & casualty_home_area_type***
  - Irrespective of gender most casualties are from urban area
- ***casualty_class & casualty_imd_decile***
  - No of casualties decrease as imd decile improves for all classes.
- ***casualty_home_area_type & casualty_imd_decile***
  - *For urban casualties decreases, For semi-urban casualties increases* as imd decile improves
- ***sex_of_casualty & casualty_imd_decile***
  - Irrespective of gender casualties decreases as we go higher the imd decile.
- ***sex_of_casualty & age_band_of_casualty***
  - Age group(26-35) has highest no of casualties irrespective of gender



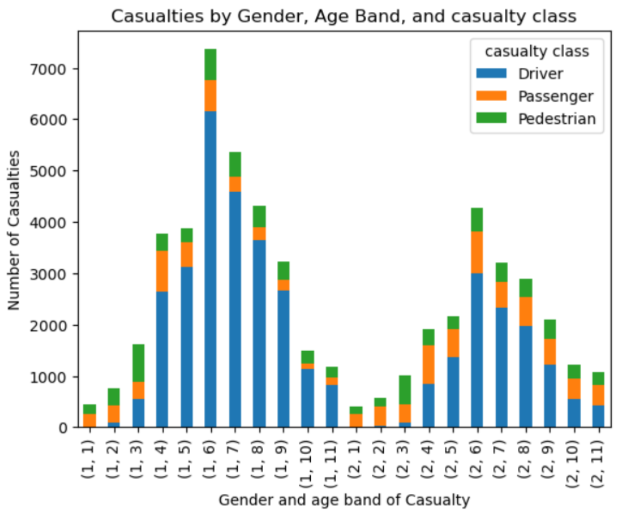


Multivariate Analysis

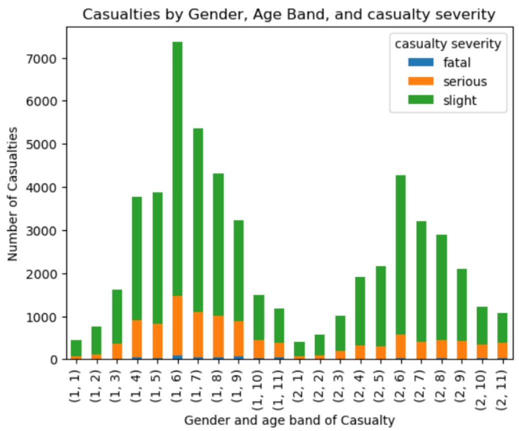
Multivariate Analysis was performed using stacked bar graph

Columns on which multivariate analysis was performed.

- ***sex_of_casualty , age_band_of_casualty & casualty_class***
  - Male and Female both have highest casualties in age group 6(26-35) with driver having the most no. in that age group



- ***sex_of_casualty ,age_band_of_casualty & casualty_severity***
  - Male and Female both have highest casualties in age group 6(26-35) with slight injury having the most no. in that age group



Distribution

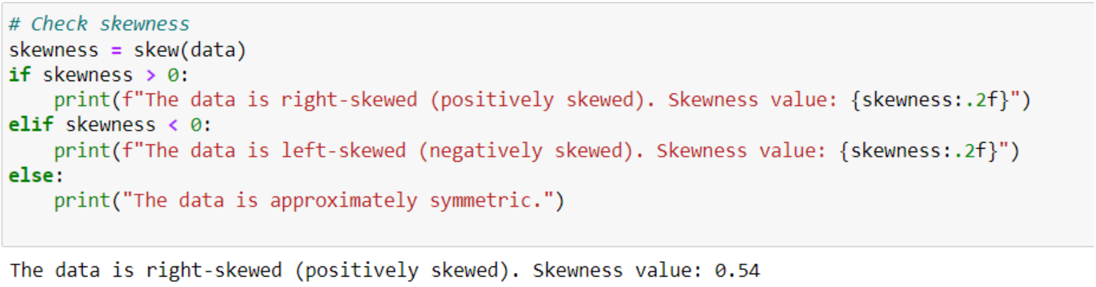
Distribution for age of casualty column was visualized using sns.histplot(d['age_of_casualty'],kde=True)

A normal distribution plot was plotted to compare it with the distribution of the age_of_casualty column .

As it can be seen that the data is not following a normal distribution.

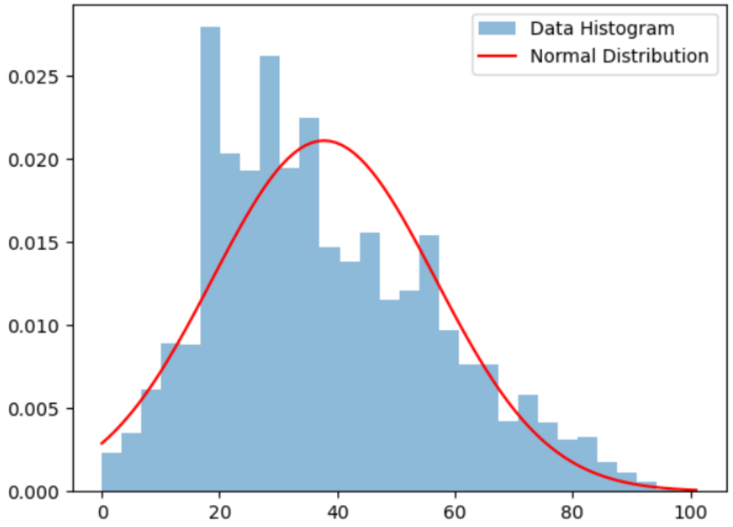
Nor is it following any other standard distributions.

Then skewedness was checked for the distribution using



```
# Check skewness
skewness = skew(data)
if skewness > 0:
    print(f"The data is right-skewed (positively skewed). Skewness value: {skewness:.2f}")
elif skewness < 0:
    print(f"The data is left-skewed (negatively skewed). Skewness value: {skewness:.2f}")
else:
    print("The data is approximately symmetric.")
```

```
The data is right-skewed (positively skewed). Skewness value: 0.54
```



The distribution was found to be right skewed.

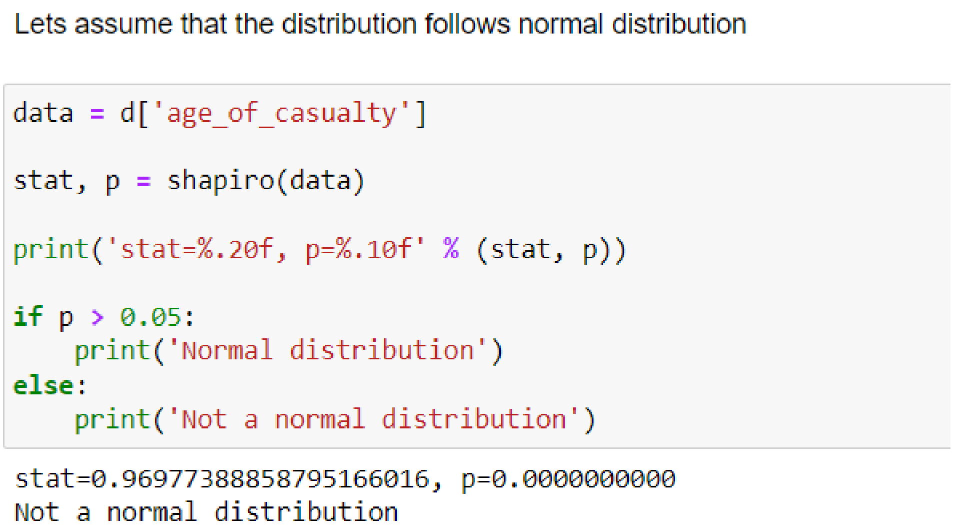
Therefore the distribution of age_of_casualty column is a **right skewed distribution**

Hypothesis Testing

Following hypothesis were developed and then tested on the data.

1.Normality test using Shapiro-Wilk Test : tests if data is normally distributed.

Normality test was conducted using Shapiro-Wilk method and the distribution was found to be not normal distribution.



Lets assume that the distribution follows normal distribution

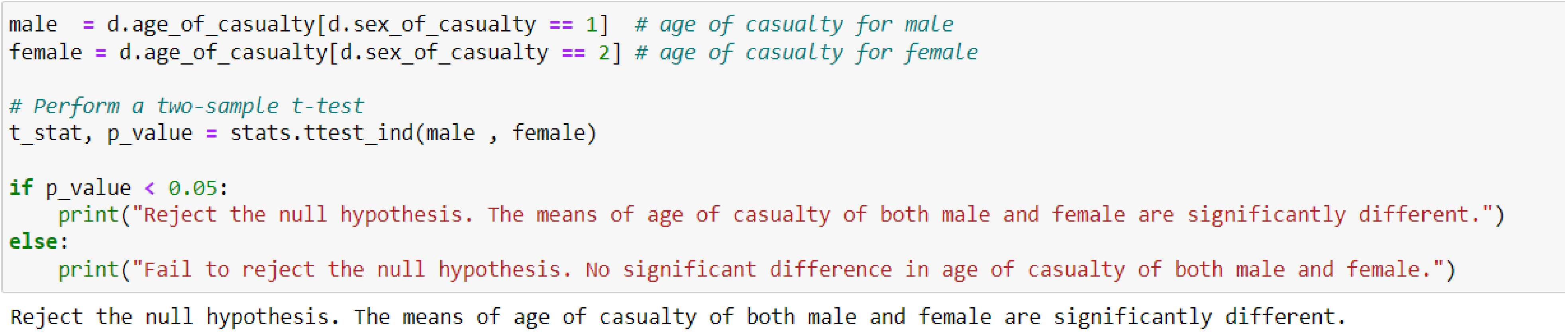
```
data = d['age_of_casualty']

stat, p = shapiro(data)

print('stat=%.20f, p=%.10f' % (stat, p))

if p > 0.05:
    print('Normal distribution')
else:
    print('Not a normal distribution')
```

```
stat=0.96977388858795166016, p=0.0000000000
Not a normal distribution
```

Hypothesis Testing

2.T Test : Comparing mean age of the casualty of male and female

Lets assume mean age of casualty of both male and female have no major difference



```
male  = d.age_of_casualty[d.sex_of_casualty == 1]  # age of casualty for male
female = d.age_of_casualty[d.sex_of_casualty == 2] # age of casualty for female

# Perform a two-sample t-test
t_stat, p_value = stats.ttest_ind(male , female)

if p_value < 0.05:
    print("Reject the null hypothesis. The means of age of casualty of both male and female are significantly different.")
else:
    print("Fail to reject the null hypothesis. No significant difference in age of casualty of both male and female.")
```

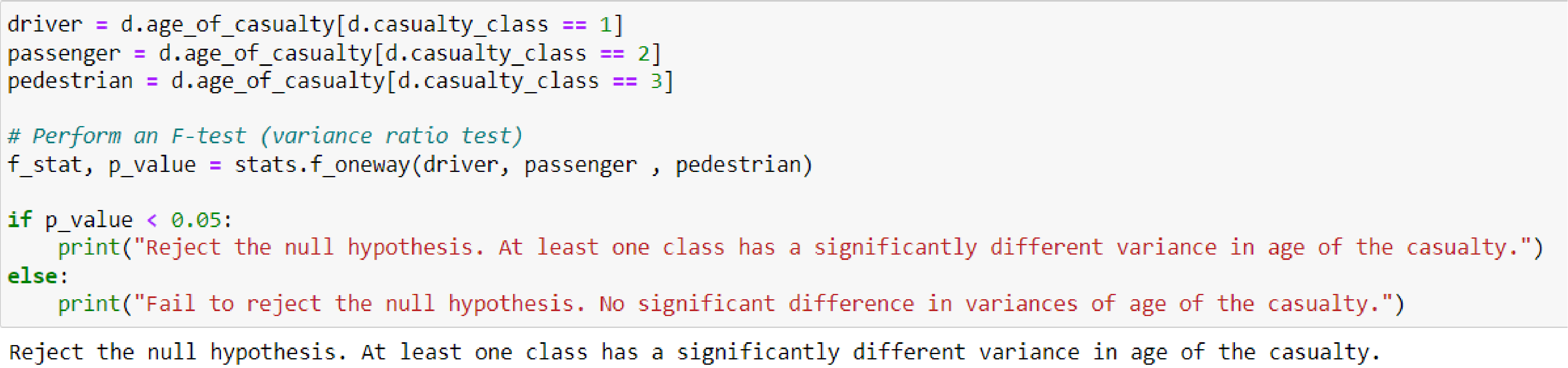
```
Reject the null hypothesis. The means of age of casualty of both male and female are significantly different.
```

the null hypothesis was rejected

Hypothesis Testing

2.F Test : Comparing variance in age of the casualty for driver, passenger and pedestrian.

Lets assume there is no significant difference in variances of age of the casualty for each group



```
driver = d.age_of_casualty[d.casualty_class == 1]
passenger = d.age_of_casualty[d.casualty_class == 2]
pedestrian = d.age_of_casualty[d.casualty_class == 3]

# Perform an F-test (variance ratio test)
f_stat, p_value = stats.f_oneway(driver, passenger , pedestrian)

if p_value < 0.05:
    print("Reject the null hypothesis. At least one class has a significantly different variance in age of the casualty.")
else:
    print("Fail to reject the null hypothesis. No significant difference in variances of age of the casualty.")
```

```
Reject the null hypothesis. At least one class has a significantly different variance in age of the casualty.
```

the null hypothesis was rejected

Finding and Insight

1. Male are more likely to be involved in a road accident.
2. People in their later 20’s and early 30’s are more likely to be involved in accidents.
3. Driver are most likely to be involved in the accidents and pedestrian are least likely.
4. 79% accident result in only a slight injury.
5. People from lower decile(or more deprived communities) will have higher chance getting in an accident that the one from higher imd decile.
6. Driver have the most chance to get any severity .
7. Pedestrian have 2nd most chance get a serious injury.
8. Passenger have 2nd most chance of getting a slight injury.
9. Driver in age group 26-35 have the highest no. of casualties in all classes.
10. Passenger in age group 16-20 have the highest no. of casualties in its own class.
11. Pedestrian in age group 11-15 have the highest no. of casualties in its own class.
12. A male have the higher chance of getting any type of injury.
13. Most driver and pedestrian caught in an accident are male.
14. Majority of passenger caught in an accident are female.
15. Irrespective of class or gender majority of people caught in an accident are from urban area .
16. IMD decile 2 have the most no .of casualties.

Conclusion

After performing Exploratory Data Analysis (EDA) on the Road Accidents dataset, I can say that it has provided valuable insights and recommendations about problems through pattern discovery, hypothesis testing, and checking of assumptions.

EDA has also helped me understand the dataset, identify trends, and find relationships between variables. It has also helped me in identifying invalid values, and anomalies in the data.

Some of the other benefits of performing EDA on the Road Accidents dataset are understanding the content, identifying trends.

EDA has helped me to identify groups that are at a higher risk of getting involved in accidents

Overall, performing EDA on the Road Accidents dataset has provided valuable insights into the factors that lead to accidents .