Data Science Report

Group - 45

Anand Samuel Gunti (201687197)

Mayur Shridhar Bhoyar (201669946)

Pavan Gorantla Venkatachalapathy (201666987)

Venkat Gopinath Polamuri (201679725)

Introduction

Road safety is an ongoing concern for drivers in any part of world and the scar it imprints on the victims and their families is unimaginable. The exponential growth of the automobile industry puts more emphasis and need to improve Road safety. Firstly, to improve we should be aware of the existing conditions and what are the major factors in the accidents and the aftermath of it.

In this Report we aim to analyse and provide insights into some of the factors based on the STATS19 dataset

AIM:

Analyse the data containing some road safety data (called STATS19) for England for the 12 months to 31/12/2019, using data and analysis to answer the following questions:

1. What patterns are there in the demographics of casualties?
2. Looking at accidents which included a killed or seriously injured (KSI) casualty, what patterns are there between the local authorities?
3. What patterns are there in pedestrians who were KSI casualties?

The Data

We have used two datasets for our analysis:

dft-road-casualty-statistics-accident-2019.csv

dft-road-casualty-statistics-casualty-2019.csv

The datasets are **publicly available** on UK government's website.

Data Quality

Data decoding has been carried out as per the Meta data file, and the requirements of the project

There are several data qualities issues in the given data set:

Format Issues

- Column **age_band_of_casualty** has issues with the wrong data entry. The data was entered as 2022-**11-15** and 2022 - **06-10,** instead of the age band **11-15 and 06-10.**
- Date time format has been changed to object data type.

Data Pre-processing

- As per meta data 8 police forces were discontinued in 2019 which were Northern, Grampian, Fife, Central, Strathclyde, Dumfries and Galloway, Lothian and Borders, Tayside. So, we have excluded these police forces because there was no data related to this in the meta data file.
- Column **speed_limit** has missing values. So, we have dropped the missing values and checked for the valid values as per meta data. In total 84 values were removed.

Data Characterization:

We have converted above mentioned data sets (csv files) into a pandas Data Frame and merged the data sets into a single dataset using the merge method in python

Description of Columns used in analysis:

These were selected based on requirements like Demographic related data, Killed or Seriously injured (KSI) data and factors that influence it.

1. **CASUALTY_SEVERITY:** Severity of the person involved
2. **SEX_OF_CASUALTY:** gender of the person involved
3. **CASUALTY_CLASS:** Describes whether the person is a passenger, driver, or pedestrian
4. **CASUALTY_TYPE:** Mode of transport the person was using.
5. **AGE_BAND_OF_CASUALTY:** Determines the age of casualty
6. **ACCIDENT_INDEX:** Serial number of accidents
7. **ACCIDENT _SEVERITY:** Describes how serious the accident was
8. **NUMBER_OF_CASUALTIES:** Number of casualties involved in an accident
9. **DAY_OF_WEEK:** The day when the accident happened
10. **ROAD_TYPE:** It specifies road type where accident had taken place
11. **JUNCTION_DETAIL:** Type of junction roundabout or crossroad etc.
12. **JUNCTION_CONTROL:** State of junction control whether automatic or controlled etc.
13. **LIGHT_CONDITIONS:** condition of light at time of accident.
14. **WEATHER_CONDITIONS:** weather condition at time of accident
15. **URBAN_OR_RURAL_AREA:** Area type where accident has occurred
16. **DID_POLICE_OFFICER_ATTEND_SCENE_OF_ACCIDENT:** Describes whether police attended the accident scene or not

VARIABLE	Type	Sample Data	Missing values
CASUALTY_SEVERITY	Categorical	Fatal, Serious, Slight	No Data Missing
SEX_OF_CASUALTY	Nominal	Male, Female	713 Values Missing
CASUALTY_CLASS	Categorical	Driver, Passenger	No Data Missing
CASUALTY_TYPE	Categorical	Cyclist, Taxi, Bus etc.,	5 Values Missing
AGE_BAND_OF_CASUALTY	Nominal	0-5, 6-10…., over 75	3234 Values Missing
ACCIDENT_INDEX	Numerical	2019010128300	No Data Missing
ACCIDENT_SEVERITY	Categorical	Fatal, Serious, Slight	No Data Missing
NUMBER_OF_CASUALTIES	Numerical	1,2,3,5 etc.,	No Data Missing
DAY_OF_WEEK	Date/time	Monday, Tuesday etc.,	No Data Missing
ROAD_TYPE	Nominal	Single, dual carriageway etc.,	No Data Missing
JUNCTION_DETAIL	Nominal	Slip road, crossroads etc.,	1 Value Missing
JUNCTION_CONTROL	Nominal	Stop sign, Auto traffic signal etc.,	62157 Values Missing
LIGHT_CONDITIONS	Nominal	Daylight, darkness (lit & unlit) etc.,	1 Value Missing
WEATHER_CONDITIONS	Nominal	Wind, snow, raining etc.,	No Data Missing
URBAN_OR_RURAL_AREA	Categorical	Urban, rural	29 unknown values

DID_POLICE_OFFICER_ATTEND_SCENE_OF_ACCIDENT	Boolean	Yes or No	9943 Values Missing

Detailed Analysis

a):- What patterns are there in the demographics of casualties?

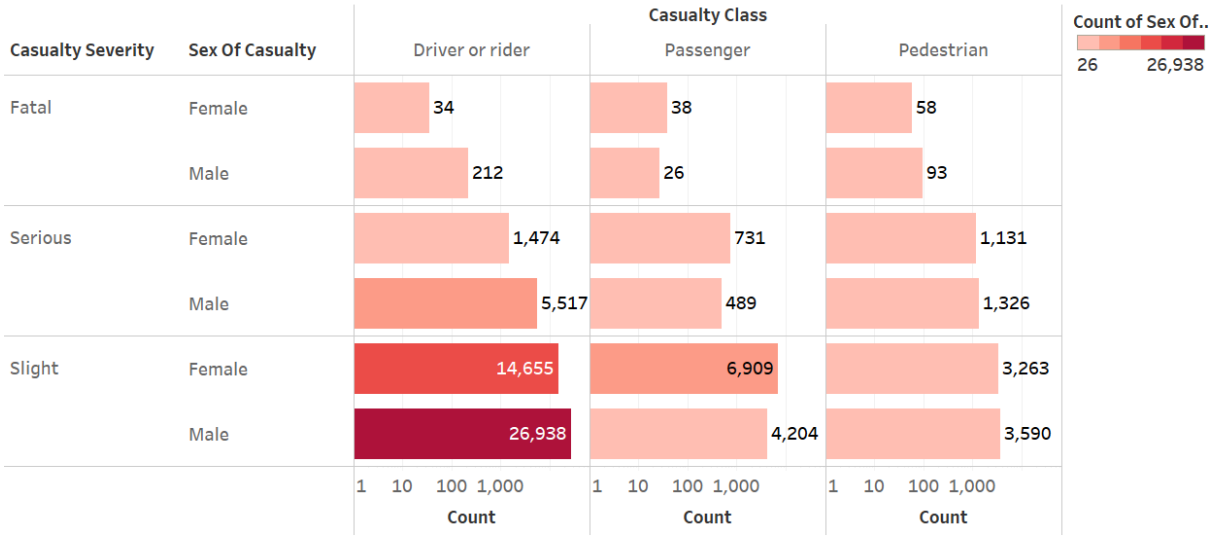

Fig. 1.1

The **Fig 1.1** is a relation between casualty severity, sex, and the casualty class of how many got injured in accidents. From **Fig : 1.1** we can say that there are maximum number of people who are driver or riders having Slight injuries with females count as 14,655 and Male count as 26,938, But in terms of Passengers we can observe that maximum number of slight injuries are happening with Females. However, In the case of Pedestrians, we can observe that both sexes are almost equally getting slightly injured.

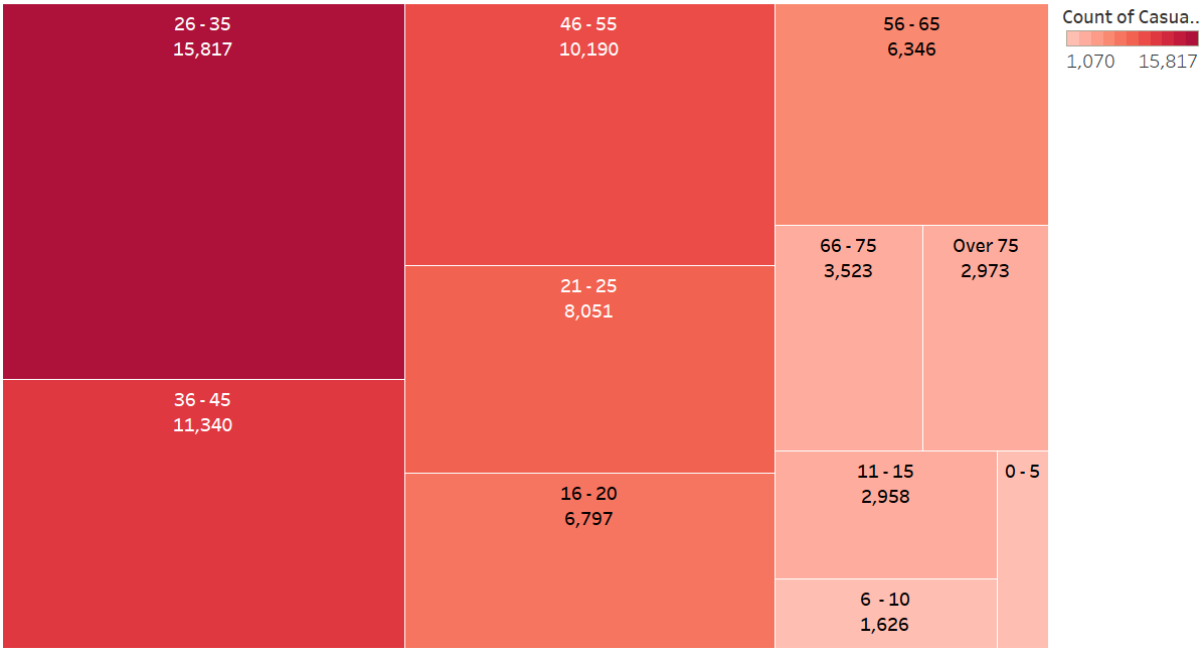

Fig. 1.2

Fig 1.2 represents the count of casualty based on the age band with respect to colour density. We can infer that the age band 26 to 35 age are having the maximum number of casualties with 15,817 getting involved in accidents. The casualties in the age band of 36 to 45 and 46 to 55 are nearly same around 11,000.

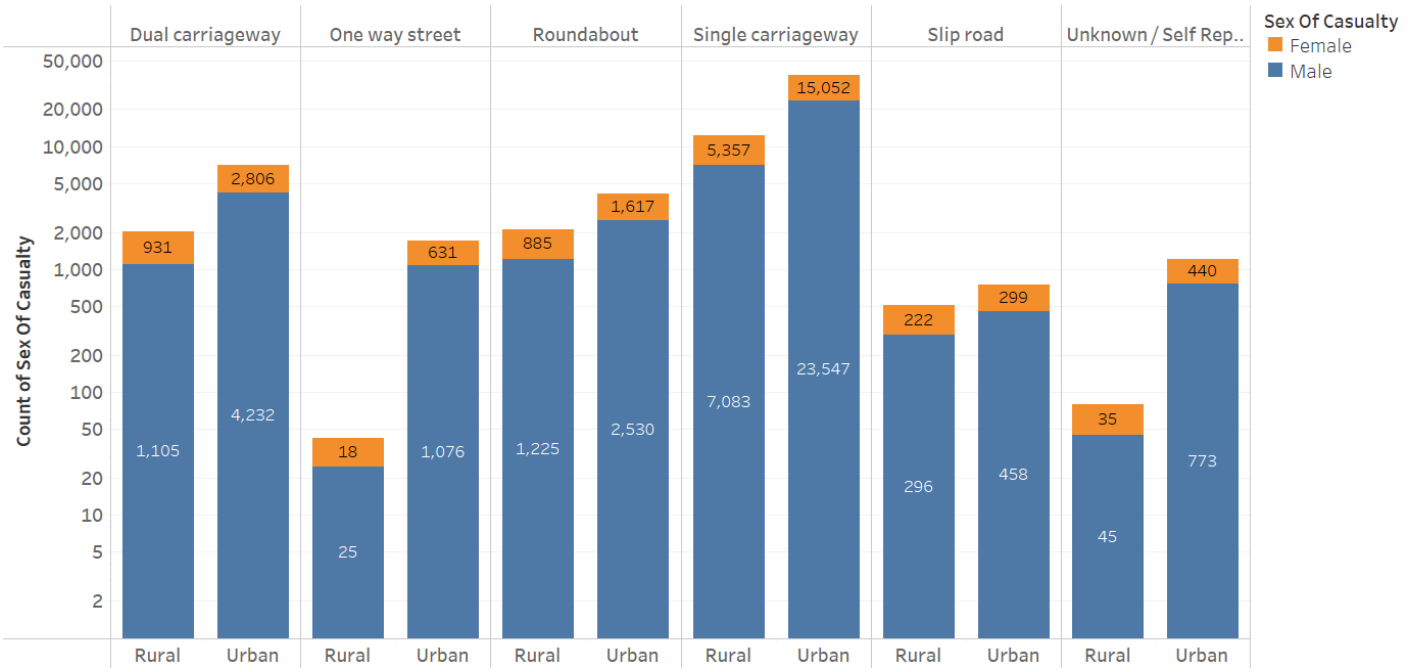

Fig. 1.3

The **Fig. 1.3** is a relation between accidents that happened on what road type based on the area, rural or urban and sex of casualty. From **Fig 1.3** we can observe that the urban Single carriageway road type is having the maximum number of casualties with around 15,000 females and 23,500 males, and in the urban Dual carriageway with maximum number of casualties with nearly 3000 female and 4000 male casualties. we can also infer that the casualties in males are more compared to females and also the data suggests that the casualties are likely to be in urban areas as compared to rural areas

b). Looking at accidents which included a killed or seriously injured (KSI) casualty, what patterns are there between the local authorities

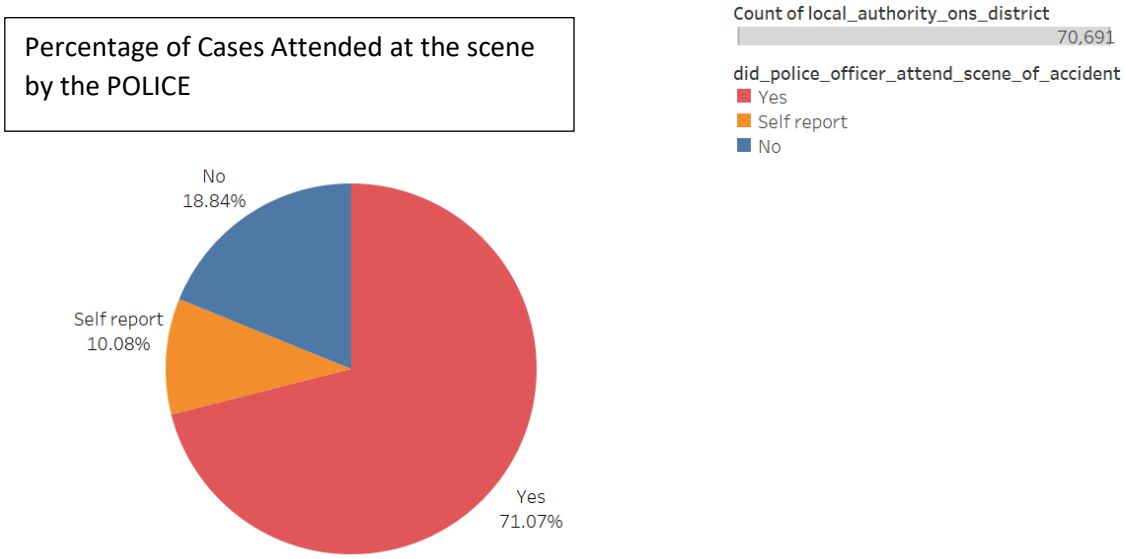

Fig 2.1: How many cases have been attended by the police?

From a total of 70691 accidents, we could depict that 71.07% of cases have been attended and 18.84% of cases have been left unattended by the police at the accident scene and moreover 10.08% of cases were self-reported. Therefore, a total of 28.94% of cases were not attended by the police at the accident scene.

“Why were the cases unattended?”

Assumption 1: Were the unattended cases occurred during the busiest period of the week?

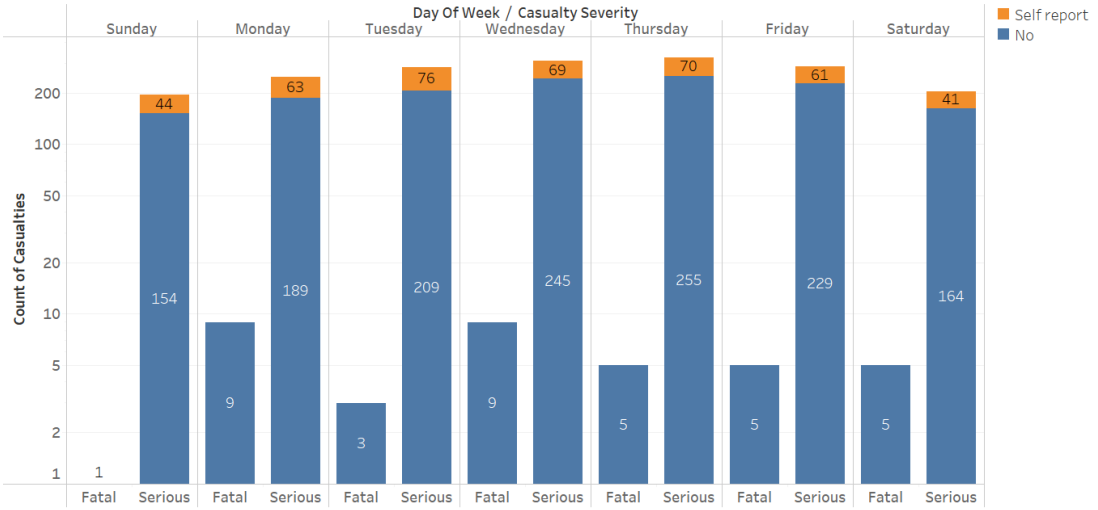

Fig 2.2

From the **Fig 2.2**, an approximate count of 5 fatal and 200 serious cases were left unattended on each day of the week. We could see that the graph across each day looks the same and there is no specific pattern associated with the unattended cases. In conclusion, the graph denies the assumption as the number unattended cases is almost uniformly distributed across the week.

Assumption 2: Was the case unattended due to a certain weather condition?

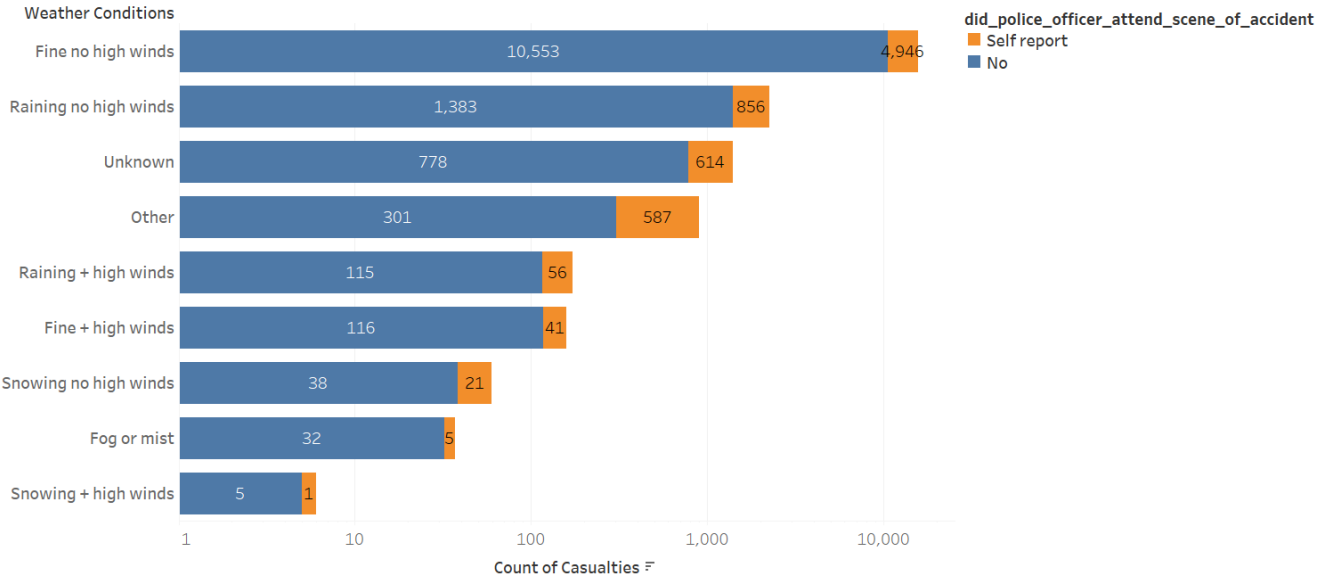

Fig 2.3

From the above bar graph, the most unattended cases were during ‘Fine no high wind’, and ‘Raining no high wind’ types of weather conditions and the least during extreme weather conditions such as snowing, foggy or raining. Even though the weather conditions were fine, there were several cases left unattended by the police force. Further, this implies that there is no correlation between unattended cases and weather condition.

c) What patterns are there in pedestrians who were KSI casualties?

For this analysis, the first question that pops up is, what is the percentage of pedestrians killed or seriously injured (KSI) casualty of total casualty class?

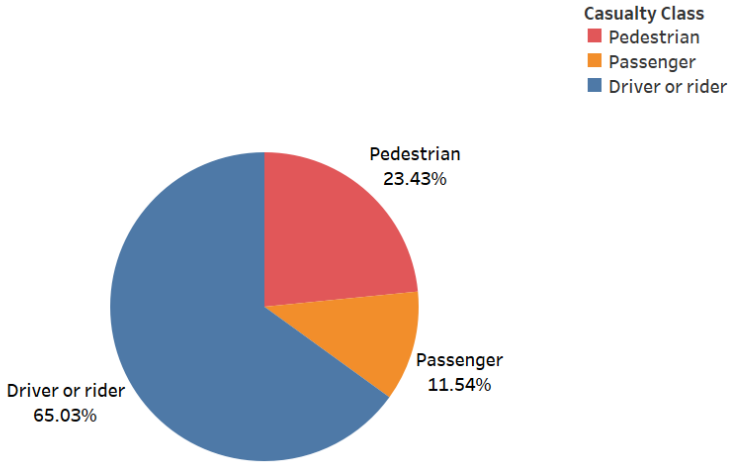

Fig. 3.1

Fig 3.1 represents a pie chart that depicts the distribution of the Casualty class. The Driver or rider class represented in blue is the class that suffered the Highest casualties with a whopping 65% and the next one being **Pedestrians** with around **24%** followed by Passengers with nearly 12%.

From **Fig 3.1** we have identified that **24%** of casualties were pedestrians and out of that what serious and fatal incidents happened are depicted in below **Fig 3.2**

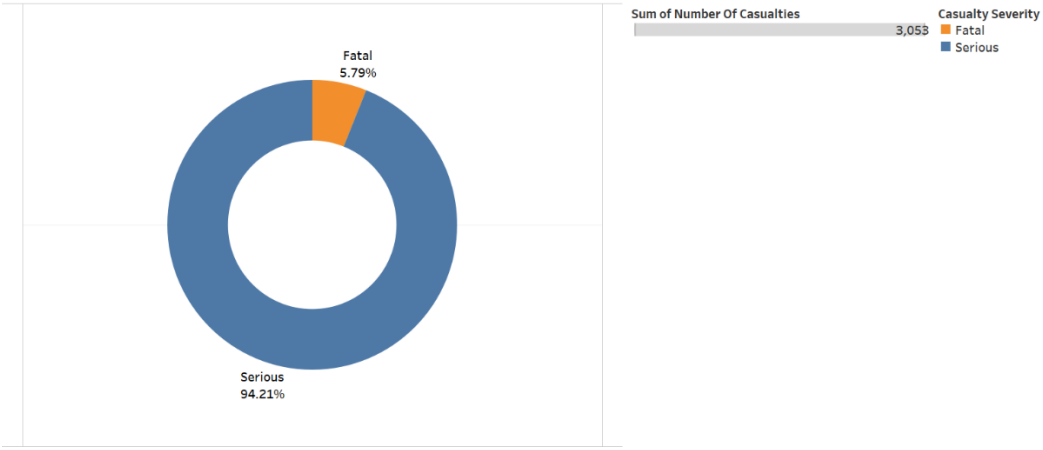

Fig. 3.2

The pie chart **Fig 3.2** represents the share of Fatal and Serious Casualties reported with Pedestrians. **94%** of reported Casualties are serious in nature and around **6%** are Fatal in nature.

Now the question arises what is the reason behind these casualties is it due to Light conditions?

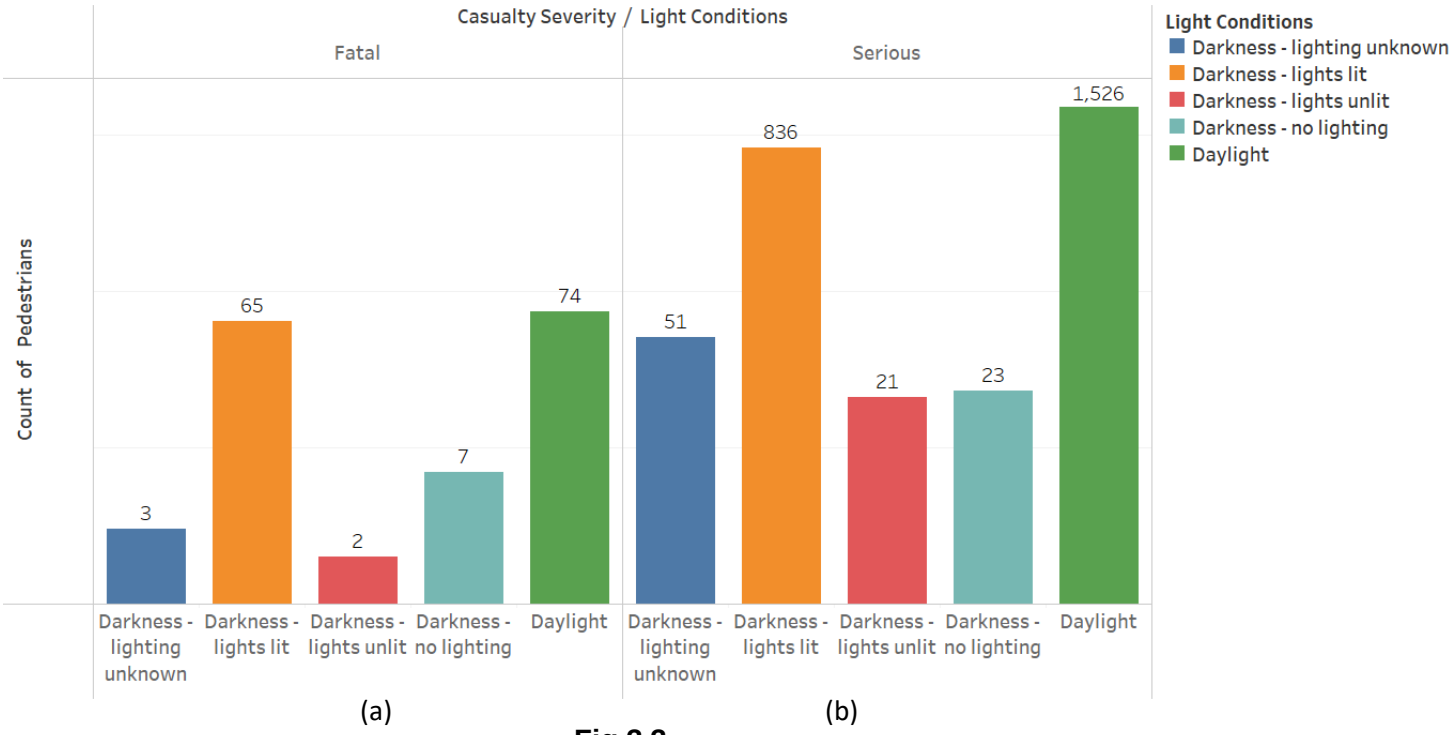

Fig 3.3

Fig 3.3 illustrates the state of light Conditions where the casualties are Serious and Fatal. The light conditions are considered based on the time of day and visibility**. From 3.3(a) and 3.3(b), 74 Fatal causalities occurred during daylight and 1526(more than 50%) serious causalities also occurred** during daylight . When the light conditions were dark and lit at night, 65 Fatal and 836 Serious casualties occurred.

From these details, we can infer that **most** of the incidents **occurred during the daylight** and the next one occurred in the darkness where the lights are lit. Only a **mere 1.8%** of total incidents have occurred during darkness and no lighting conditions.

From the above findings, we can infer that **most casualties occurred in daylight**. The reason behind this is unknown or maybe is it a case of negligence, which is to be investigated.

The final question is that in which junctions these incidents occurred and what controls they have at these junctions

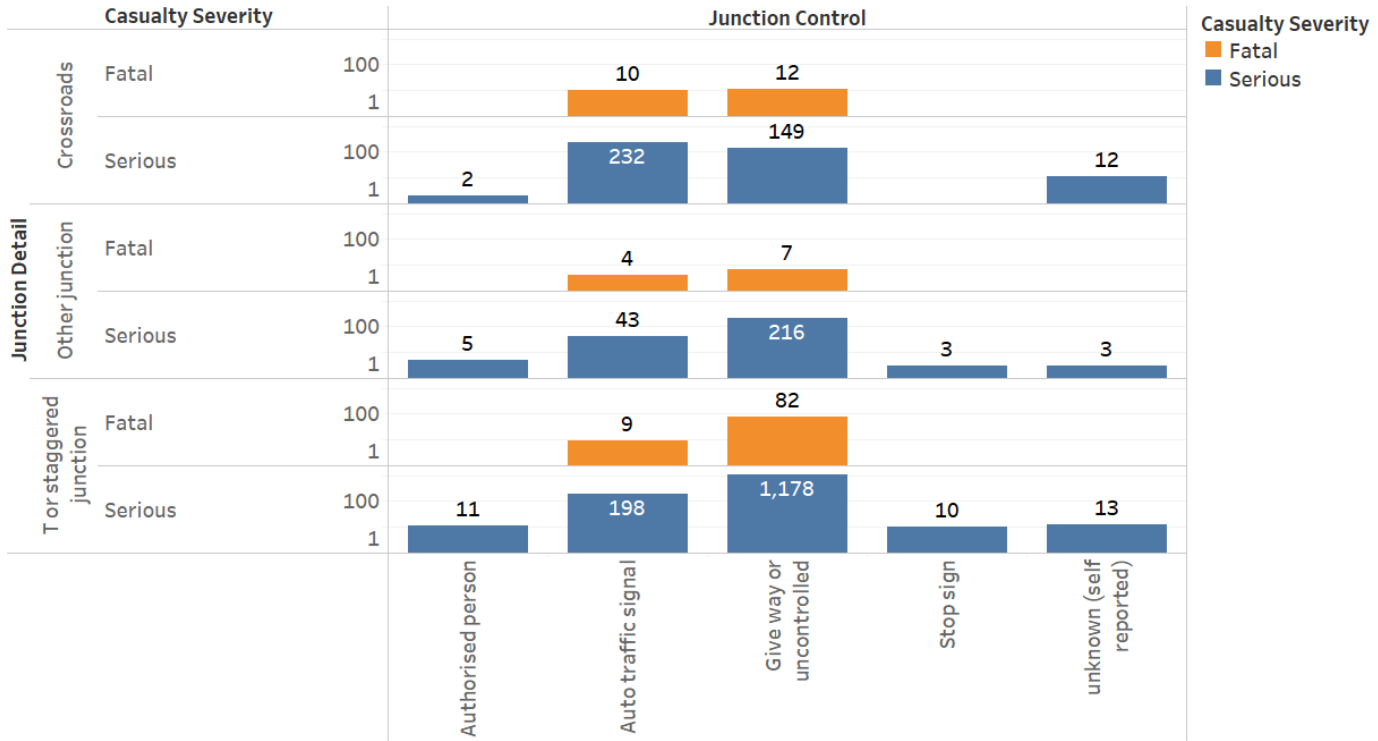

Fig 3.4

Fig 3.4 depicts the Casualty severity (Fatal or Serious) based on Junction detail and junction control.

This suggests that maximum casualties have occurred at T or staggered junctions where the junction control is a Give way or uncontrolled.

We can also infer that the Fatal causalities in pedestrians were minimum where there is Control and whenever there is an authorized person or stop sign at the junction control the number of causalities was just 79 serious casualties and Zero Fatal casualties also strengthen inference. Even though, there is an Auto traffic signal controller there were considerable serious casualties that occurred

Based on the above findings, we can safely say that the more you exert control at a junction the lesser the probability of causalities in the case of pedestrians.

Conclusion:

The raw data obtained were merged and cleansed. We have sliced the required columns according to tasks required for the analysis.

Through the casualty demographics, we could infer that 65% driver or rider of both the genders and people around the age band of 26-45 are more vulnerable to accidents of all severity types.

The above assumptions used in the analysis to study the behaviour of authorities illustrates that the weather conditions or the busiest period of the week were not the reason for the police not appearing at the accident scene. There is no additional data to prove that the police had any other valid reason for their absence.

The study on pedestrian casualty reveals that 94% of accidents involving a pedestrian were serious and mostly have occurred during the day time and at the T junction or crossroads.

There is a need to conduct further investigation regarding the cause of pedestrian casualties during day time and also deploy manpower at junction controls to reduce the casualties as the report suggest that when there is an authorised person at the junction control fatal casualties were zero