IMAT5322 BIG DATA ANALYTICS

ASSESSMENT 2

EXPLORING ACCIDENTS CAUSED BY DIFFERENT VEHICLE TYPES

Done By

Kaushik Chandrasekhar

P2656623

ABSTRACT

In this report I am doing an analysis on the Accident and Vehicle data from 2018 from <https://data.gov.uk/dataset/road-accidents-safety-data> . Initially the data is explored for Total observations, Total columns and missing values in each of the columns. Then the columns not used are dropped. Then I am using a combination of SparkSQL and Pixiedust to explore the datasets. Once the basic exploration is complete, I am doing an analysis on accidents happening with each vehicle type. This analysis is done based on 2 regions: 1) Urban and 2) Rural. We are able to see how different types of vehicles having different effects on the 2 regions.

**Contents**

1.INTRODUCTION…………………………………………………………………………….4

2.EXPLORING THE DATASETS……………………………………………………………5

3.CLEANING THE DATA…………………………………………………………………….6

4.DATA EXPLORING AND ANALYSIS………………………………………………….8

4.1 Cars…………………………………………………………………………………………14

4.2 Pedal Cycle……………………………………………………………………………..16

4.3 Motorcycles……………………………………………………………………………18

4.4 Taxi/Private hire car……………………………………………………………….20

4.5 Minibus (8 - 16 passenger seats)…………………………………………….22

4.6 Bus or coach (17 or more pass seats)………………………………………24

4.7 Agricultural vehicle…………………………………………………………….26

4.8 Van / Goods 3.5 tonnes mgw or under………………………………28

4.9 Goods over 3.5t. and under 7.5t……………………………………………..30

4.10 Goods 7.5 tonnes mgw and over……………………………………..32

4.11 Goods vehicle - unknown weight…………………………………….34

5.CONCLUSION……………………………………………………………………………….36

6.REFERENCES………………………………………………………………………………..37

1.INTRODUCTION

The datasets I am working with in this assessment are the 2018 Accident and Vehicle datasets. There are 122635 records in 32 columns and 226409 records in 23 columns in the 2 datasets, respectively. The following variables are used:

Accidents data:

Accident\_Index,

Longitude,

Latitude,

Accident\_Severity,

Number\_of\_Vehicles,

Number\_of\_Casualties,

Date,

Day\_of\_Week,

Time,

Local\_Authority,

Local\_Authority\_(Highway),

Road\_Type,

Speed\_limit,

Light\_Condition,

Weather\_Conditions,

Road\_Surface\_Conditions,

Urban\_or\_Rural\_Area

Police\_force

Vehicles data:

Accident\_Index

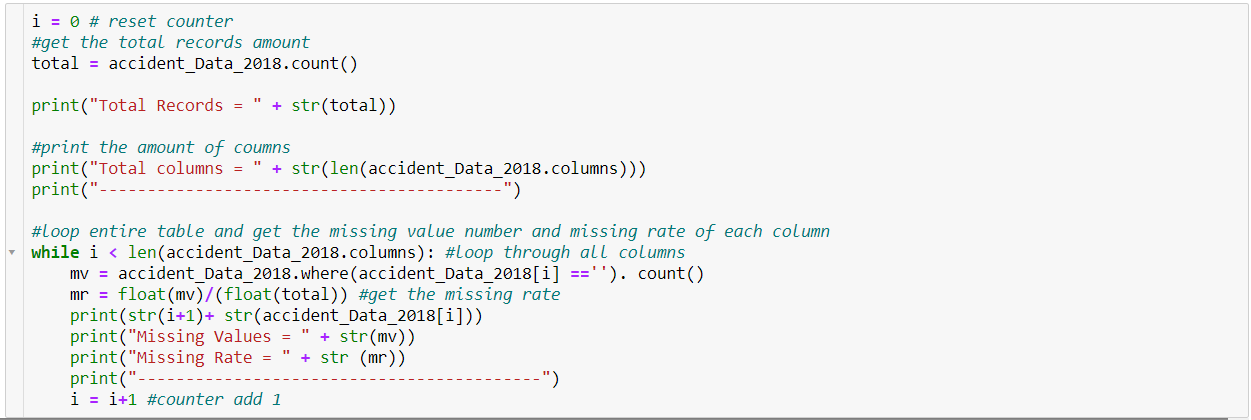
Vehicle\_Type

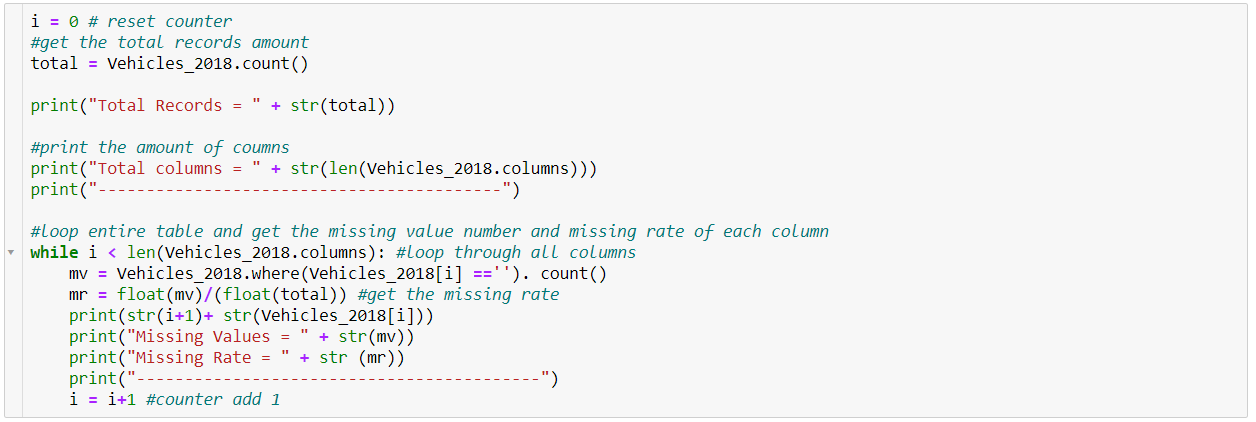
Sex\_of\_Driver

Age\_of\_Driver

With the combination of above variables, I will be performing a series of analysis using SparkSQL and Pixiedust.

2.EXPLORING THE DATASETS

The above code outputs total observations in the accident\_Data\_2018 to be 122635 with 32 columns.

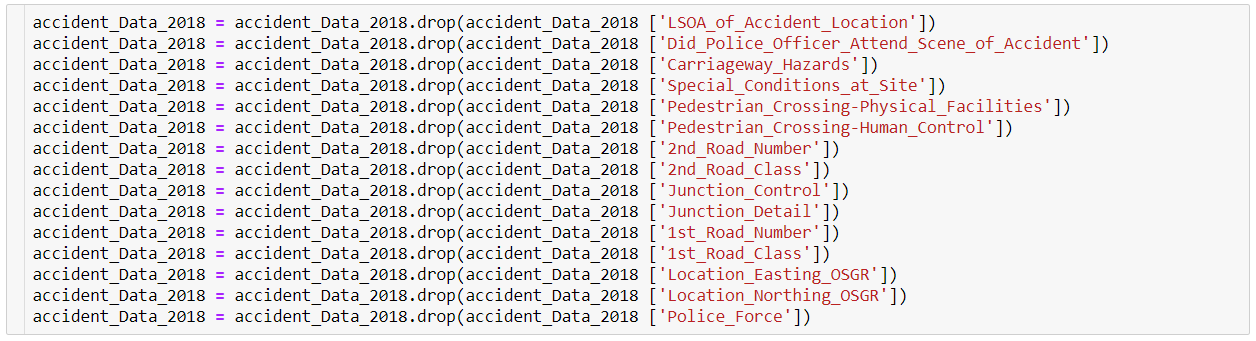


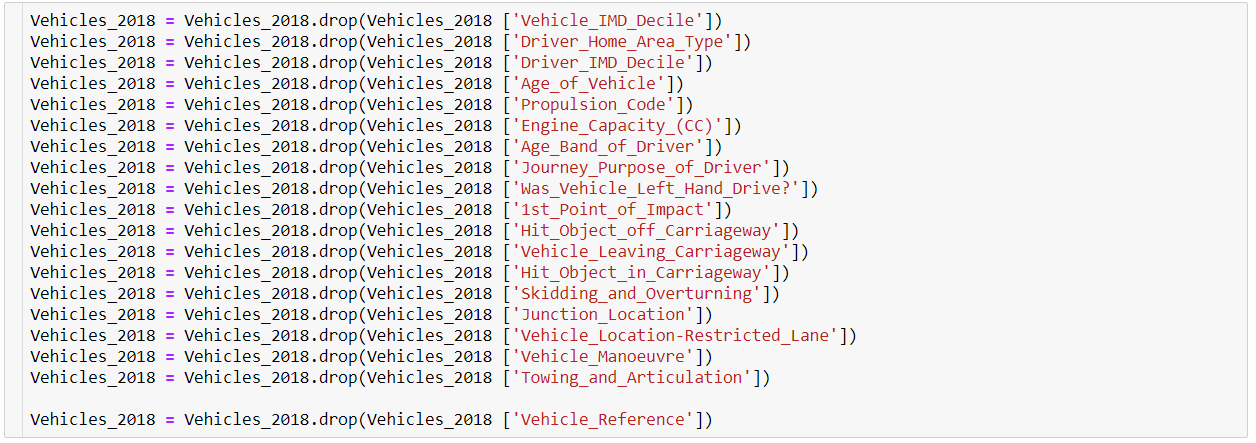
The above code outputs total observations in the Vehicles\_2018 to be 226409 with 23 columns.

There were 0 missing values in all of the variables.

3.CLEANING THE DATA

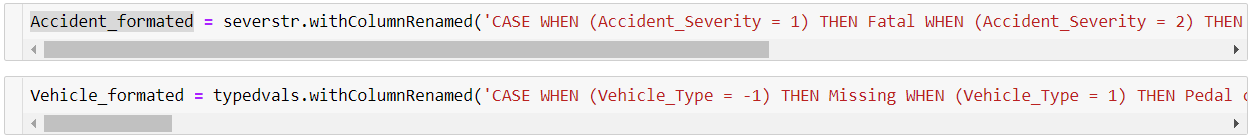
I am dropping the rest of the variables with the below code:





These variables are not used for any analysis in this report. Thus, to cut down the sheer number of variables and to reduce confusion the unused variables in the 2 data frames are being dropped.





I am assigning the designated values Vehicle\_Type and Accident\_Severity to variables Vehicle\_Type\_str and Accident\_Severity\_str and storing in a new data frames Vehicle\_formated and Accident\_formated respectively.

4. DATA EXPLORING AND ANALYSIS

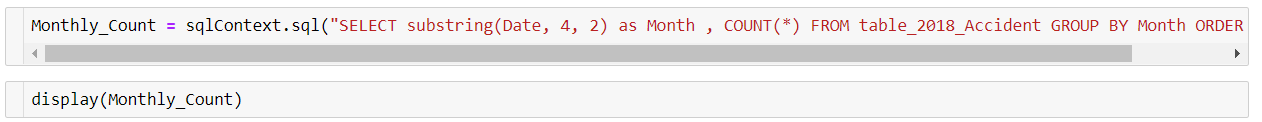
I am using a combination of SparkSQL and pixie dust to display bar graph of the data and do the analysis.

Spark SQL is a structured data processing Spark module. It can be used as a distributed SQL query engine and provides a programming abstraction called Data Frames. It also has a strong link to the rest of the Spark ecosystem (e.g., integrating SQL query processing with machine learning).

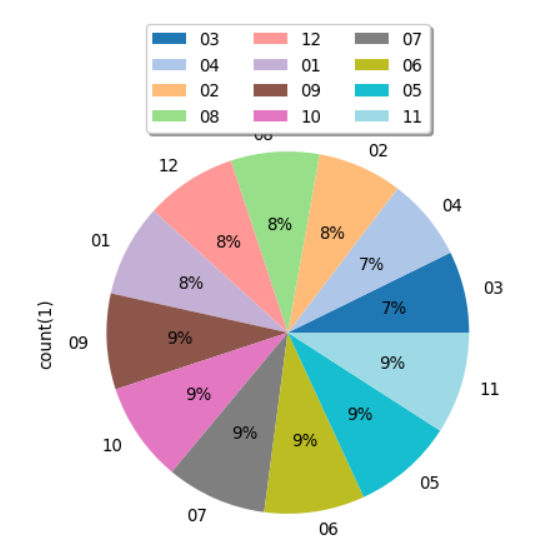
I am registering a Table table\_2018\_Accident for using SparkSQL queries from Accident\_formated.



The Below code is used to create a new Monthly\_Count data frame that has the count of accidents of each month in 2018 using “substring” to get the month from the Date variable.

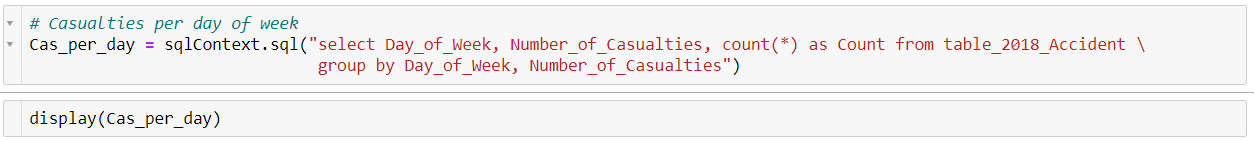


OUTPUT:

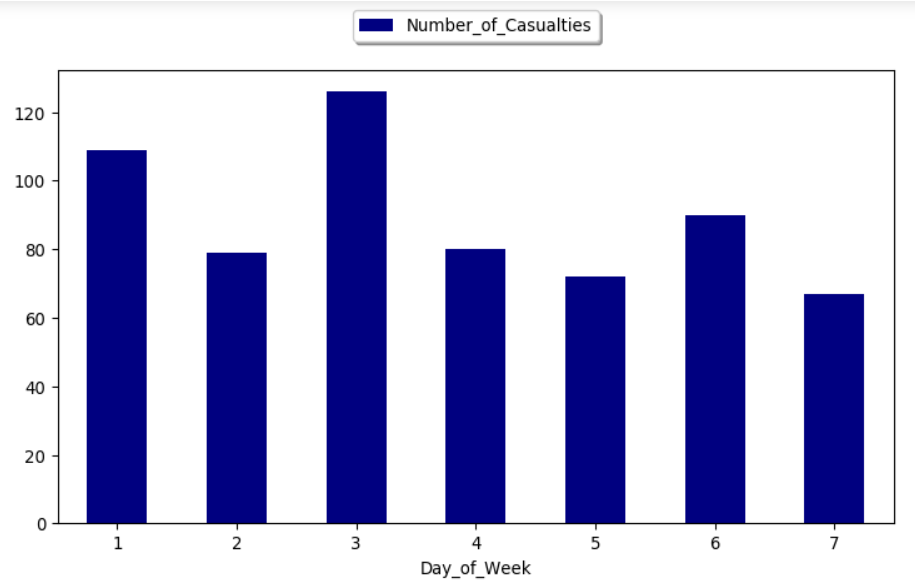


From the above pie chart we can see that Months 05, 06, 07, 09, 10, 11 are the months with the highest number of accidents with 9% each, Months 01, 02, 08, 12 comes next with 8% accidents and finally Months 03, 04 has the lowest accidents with 7% each.

The below code is used to declare a data frame Cas\_per\_day to get the number of casualties in each day of a week with SparkSQL



OUTPUT:



From the above bar chart, we can make out that Day 3 of a week has the highest number of casualties followed by day 1. Day 2 and 4 has similar number of casualties and day 5 has less casualties to day 4. The lowest number of casualties has been recorded on day 7. The highest number of casualties per day has gone over 120 in the 3rd day and the lowest number casualties is just above 60 in day 7 which represents Sunday. The Number of casualties in the weekends are comparatively lower than on working days. This could be because on the weekends there is usually less traffic and rush compared to the weekdays.

The following record creates a new data frame Accident\_Severity\_1 which has the different accident severity (fatal, Serious & Slight)

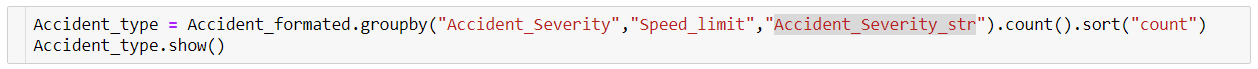


OUTPUT:



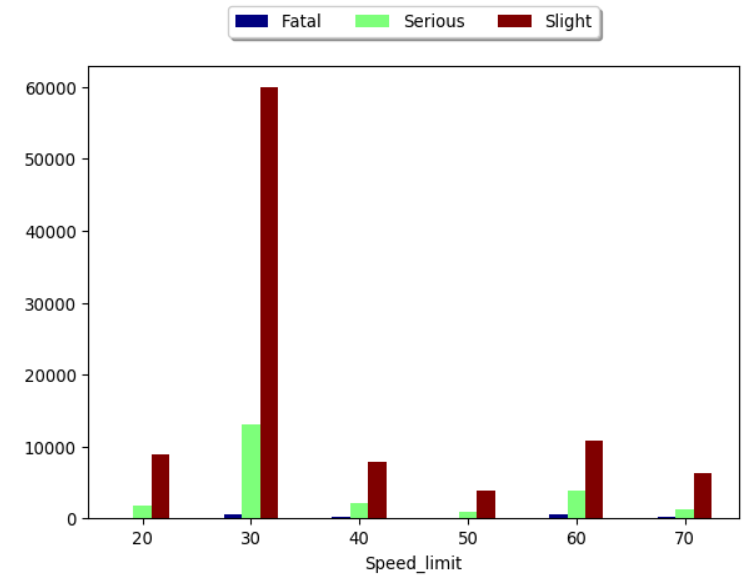
From the output table we can say that 1671 cases where fatal 23165 cases were serious, and 97799 cases were slight. So, in all of 2018 there were 1671 fatal accidents.

A new data frame Accident\_type is created from Accident\_formated with Accident\_Severity and Speed\_limit, Accident\_Severity\_str and count. This data frame is used to plot a graph between “Speed\_limit” and Accident\_Severity.

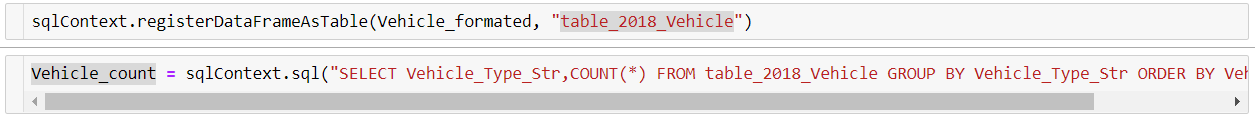




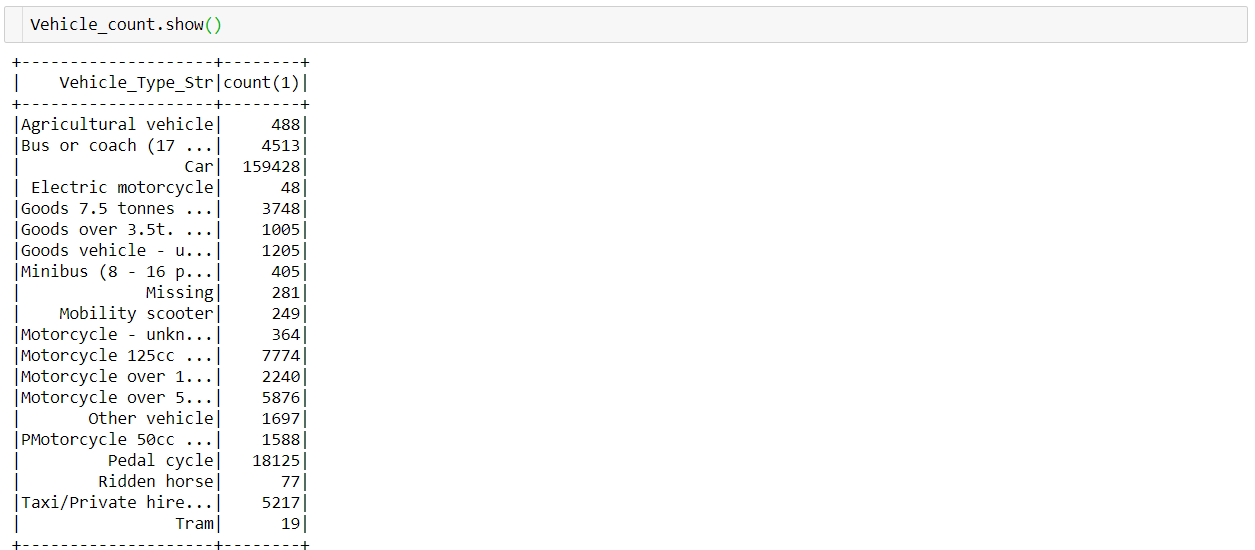
OUTPUT:



From the above graph we can determine that in Speed\_limit 30 most of the accidents occur. Fatal accidents have occurred in Speed\_limit 30, 40, 60 and 70. All the Serious accidents are equally distributed in all the Speed\_limit except 30 which has the highest. In Speed\_Limit 20 there are no Fatal accidents this could be because people are more aware of their surroundings while driving slowly. At Speed\_limit 30 we can see the highest number of accidents in which Slight accidents have the highest count followed by Serious Severity. At Speed\_limit 40 there is significant amount of Slight Severity accidents followed by Serious and few Fatal accidents. In Speed\_limit 50 there are comparatively a smaller number of accidents and from the graph we can see that there are no Fatal accidents in this Speed\_limit. At Speed\_limit 60 we can see a significant number of Slight accidents followed by Serious and a considerable number of Fatal accidents.



With the above code I am registering a new table table\_2018\_Vehicle from Vehicle\_formated to do some analysis on the vehicle data.



The Vehicle\_count data frame has the observations seen above. The count shows the number of accidents occurred for each Vehicle\_Type\_str. When I display Vehicle\_count in pixiedust I get the following result.

Agricultural vehicle - 488

Bus or coach (17 or more pass seats) - 4513

Car - 159428

Electric motorcycle - 48

Goods 7.5 tonnes mgw and over - 3748

Goods over 3.5t. and under 7.5t - 1005

Goods vehicle - unknown weight - 1205

Minibus (8 - 16 passenger seats) - 405

Missing - 281

Mobility scooter - 249

Motorcycle - unknown cc - 364

Motorcycle 125cc and under - 7774

Motorcycle over 125cc and up to 500cc - 2240

Motorcycle over 500cc - 5876

Other vehicle - 1697

PMotorcycle 50cc and under - 1588

Pedal cycle - 18125

Ridden horse - 77

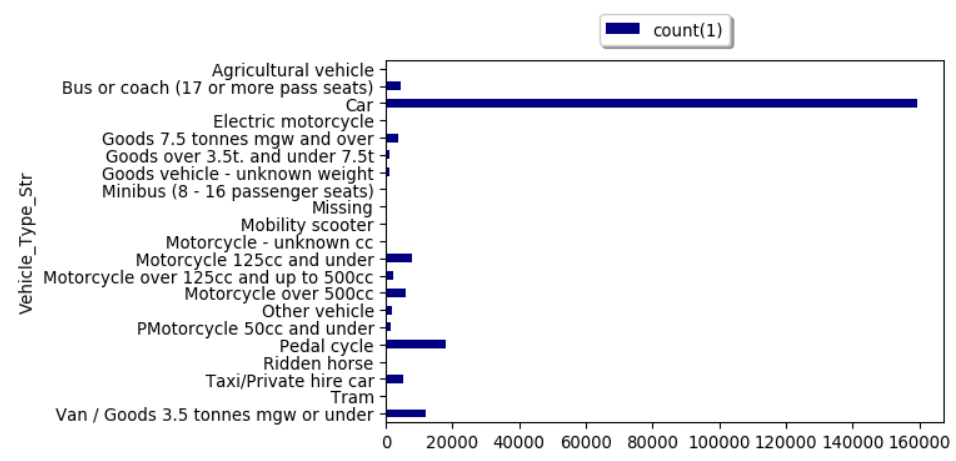
Taxi/Private hire car - 5217

Tram - 19

Van / Goods 3.5 tonnes mgw or under – 12062

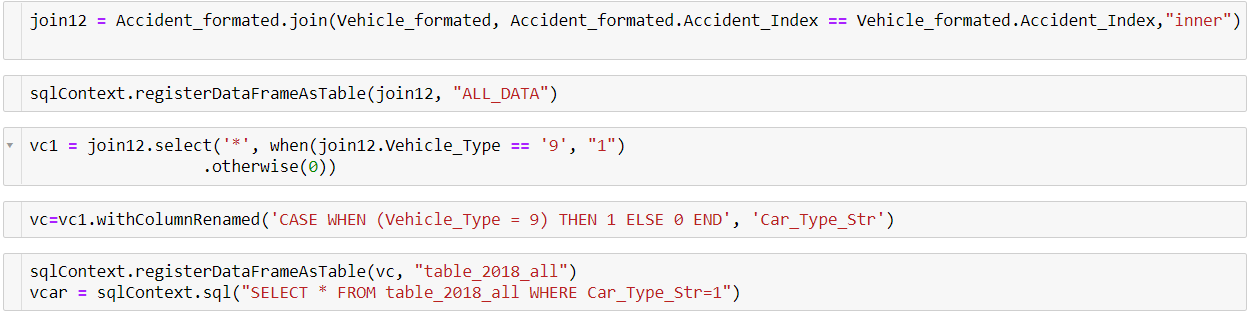
The above results show the Total number of accidents caused by the each of the Vehicle\_type. From the above data we can see that the greatest number of accidents is caused by “Cars” followed by “Pedal cycle” then comes “Van / Goods 3.5 tonnes mgw or under”. The least number of accidents is caused by “Tram”2 with a number of 19.

From the above data we can conclude that even though a “Pedal cycle’ does not go in high speed it has a high chance of meeting an accident.

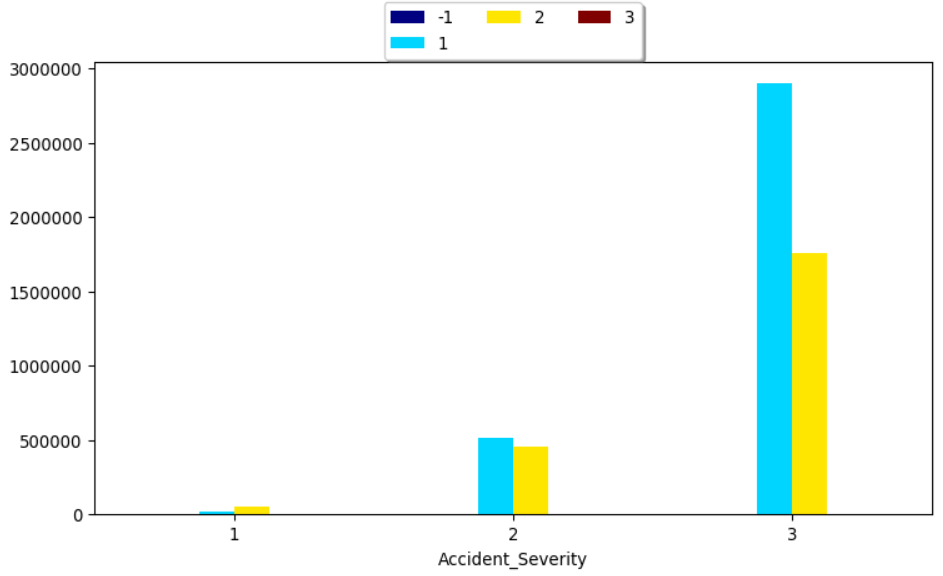


The above graph has been plotted using the Vehicle\_count data frame from which we can see the number of accidents caused by “Cars” is the highest. There are values in the fields where there is no bar, it’s not displayed as they are very small values compared to the count in “Cars”. The accidents caused by all class of motorcycles also add up to a significant amount. The total accidents occurred in 2 wheeled motorcycles adds up to 18,091 which is a significant number. The accidents occurring in pedal cycles is also very high at 18125. This number is even greater than the number in Motorcycles combined. This could be the result of improper training received for the people using pedal cycles. To use a Motorcycle the user must have a license which can only be obtained after completing a set of training. Thus, Motorcyclists tend to be more aware while riding compared to the pedal cycle counterparts. There are also 281 missing values that are not represented in the in the above graph. In the below analysis we are going to focus on the accidents caused by different categories of vehicles with the accidents greater than 100 in 2018.

**4.1 Cars**



In the code above, I am creating a new data frame join12 by merging Accident\_formated and Vehicle\_formated data frames. From there a new table ALL\_DATA is created for further analysis. Also, a data dram vc is created with a field Car\_Type\_str to flag all the data with Vehicle\_Type as 9 (Cars) as 1. This data frame is then registered as a table table\_2018\_all and the used to create another data frame vcar which contains all the accident data caused by “Cars”.

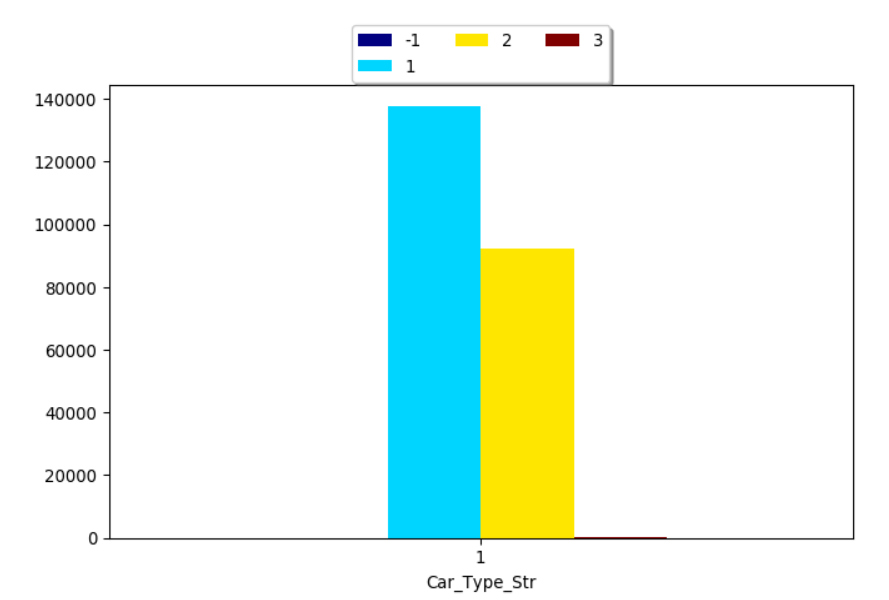


The above graph represents the Accident\_Severity for accidents caused in Urban (represented by 1 (CYAN)) and Rural (represented by 2 (Yellow)). Here 1 represents Fatal accidents, 2 represents Serious accidents and 3 represents Slight accidents.

Here we can see that the total Fatalities is more in the Rural region compared to the Urban region. This is probably because of the availability and accessibility of health care services in the Urban region. In Rural regions even if there is an availability of health care services, they will be remote and will not be accessible easily. Thus, there is an increase in fatalities in the Rural region.

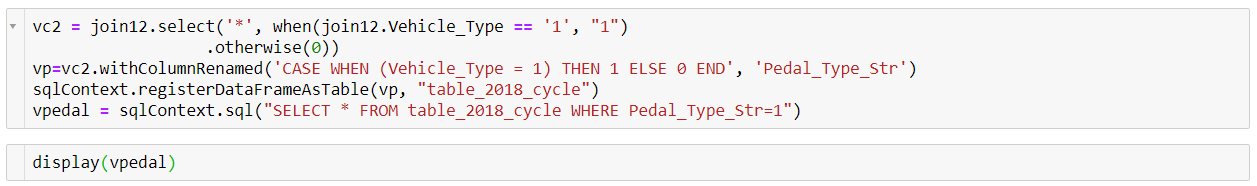
When we consider Serious accidents, we can see that the accidents in the Urban region tend to be higher than that of the Rural region. This is probably because the Number of cars in the Urban Region is more than the once found in the Rural region. Still there is no marginal difference between the two.

In the case of Slight accidents, we can clearly see that there is a considerable number of accidents in the Urban region compared to the Rural. This could be because there are lots of cars in the Urban region and chances of getting into an accident.

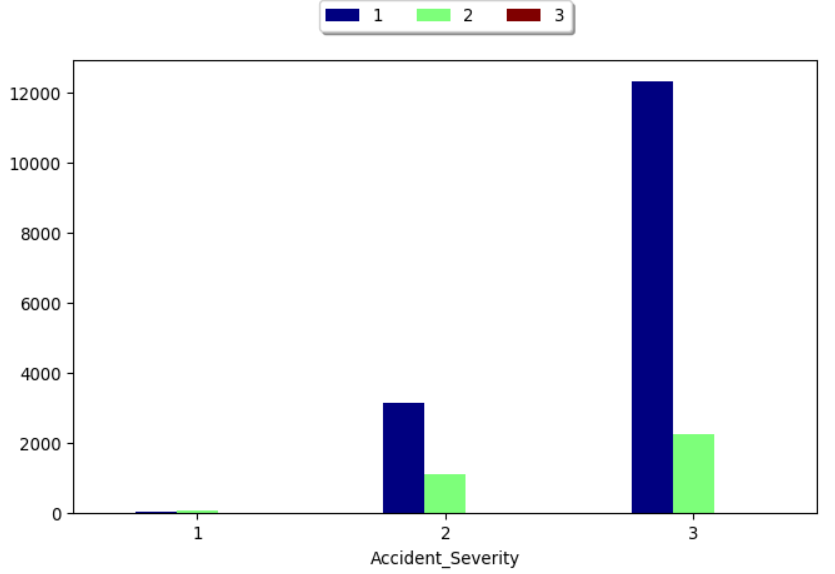


The above graph represents the total number of casualties occurred where a car is involved. The Cyan bar represents the casualties in Urban region and the Yellow bar represents the casualties in Rural region. From the above graph we can see that the is a greater number of casualties in the Urban Region compared to the Rural. The cause of this could be the sheer number of cars in the Urban Region and the rush caused by these cars.

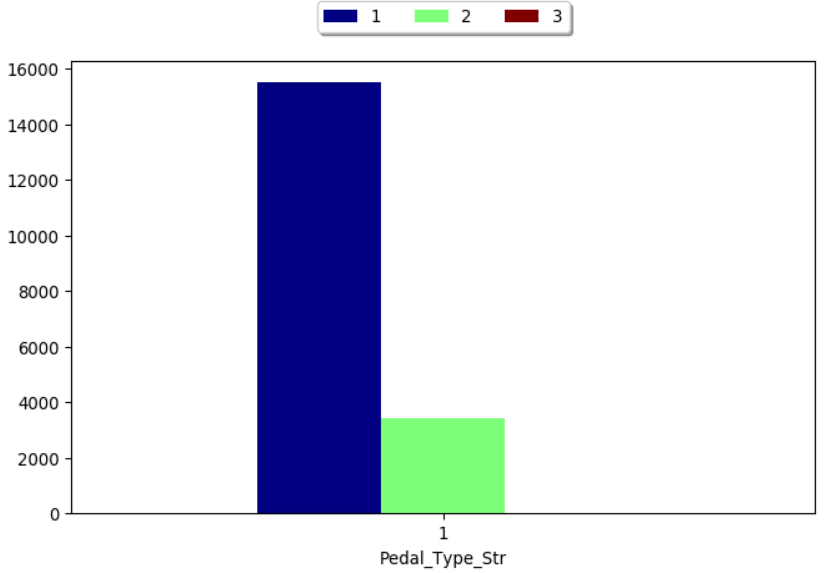
**4.2 Pedal Cycle**



The above code is used to take all the accidents caused in pedal cycles form the join12 data frame. A variable Pedal\_Type\_str is created to make a new data frame vpedal which consists of all the accidents involving a pedal cycle.

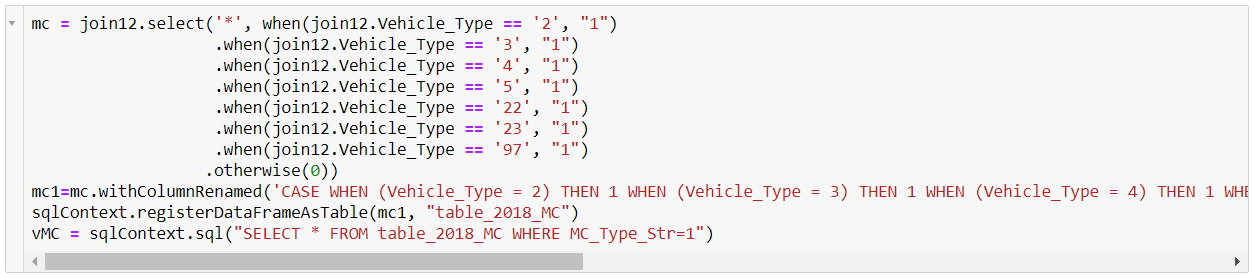


The above graph represents the Accident\_Severity and number of casualties involved in an accident with pedal cycle. Here the Blue bar represents the Urban region, and the Green bar represents the Rural Region. From the graph we can see the is a very low fatality rate (which is represented by 1) for pedal cycles. The Serious (Which is represented by 2) accidents are high in Urban region compared to rural. The Slight accidents (which is represented by 3) is very high in Urban region. With the large number of people and congested roads could be the reason for this. In Urban region its easy to get in an accident with a pedal cycle because of the crowd.

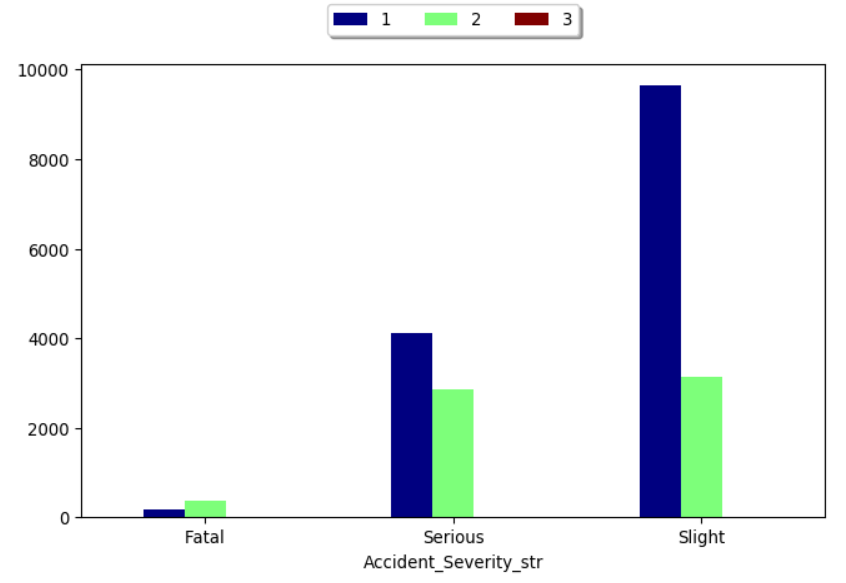


The above graph represents the number of casualties in Urban region and Rural Region using pedal cycles. The Blue bar represents the Number of casualties in Urban region which is near to 16000 and the Green bar represents the casualties in Rural region which is just above 2000. From the above graph we can clearly say that the greatest number of accidents in Pedal Cycles occur in the Urban region.

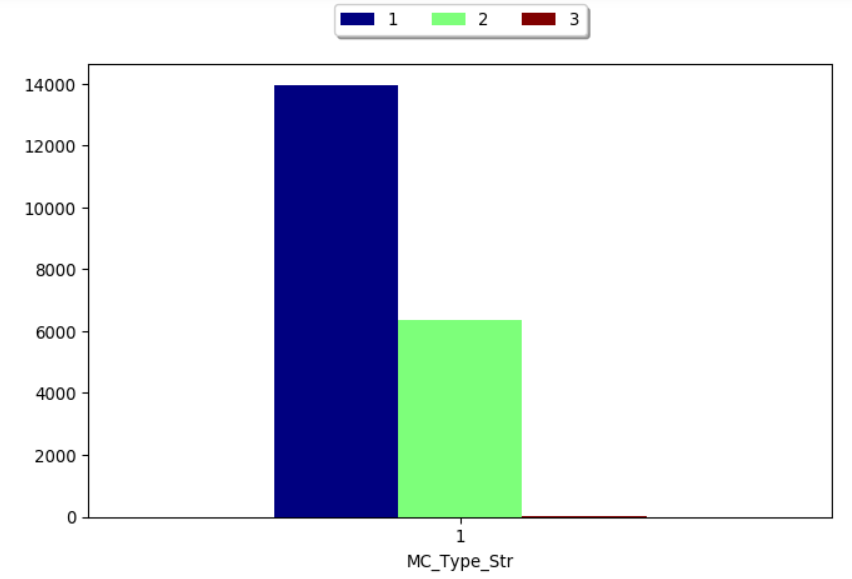
**4.3 Motorcycles**



With the above code I am using all the motorcycle data and creating a new data frame vMC with the help of SparkSQL.

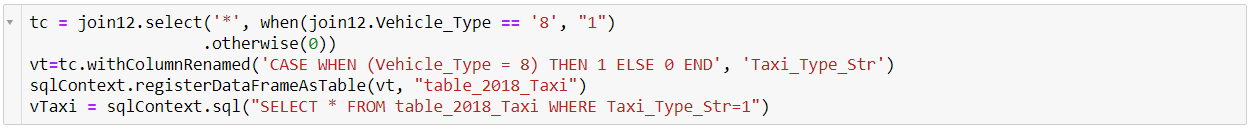


When we display the vMC with Accident\_Severity and number of casualties we get the above graph. From the above graph we can see that there is a considerable number of fatalities in the case of Motorcycles. Even for the cars data frame the Fatalities were not as great as the once in the Motorcycles. This could be because there is little protection in motorcycles compared to cars. The Rural region has the greatest number of fatalities compared to the Urban region. The number of Serious accidents is greater in the Urban region compared to the Rural region. This could be because of the number of vehicles in the Urban region compared to the Rural region. Slight accidents in Motorcycles are also great in the Urban region.

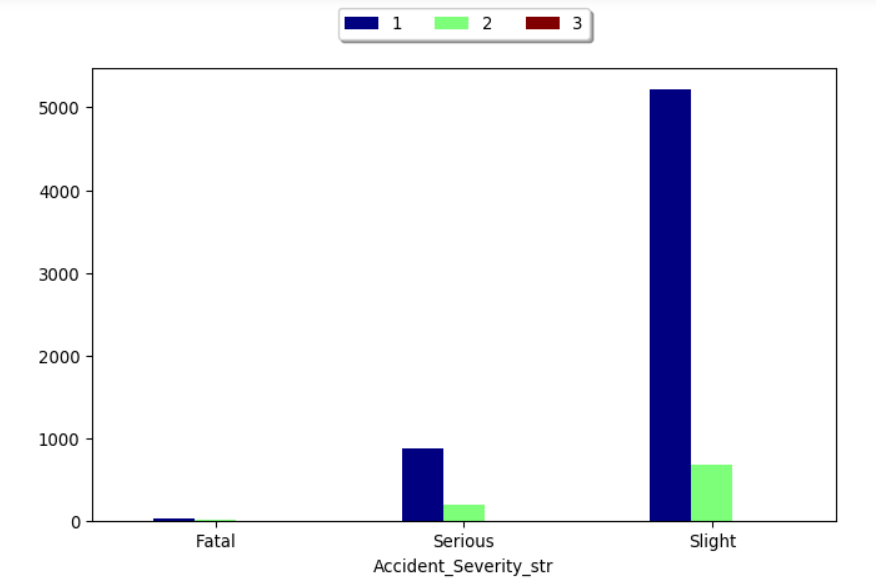


From the above graph we can conclude that the number of casualties in the Urban Region (represented by the blue bar) is greater compared to the Rural Region. The high volume of vehicles and people could be the cause of this.

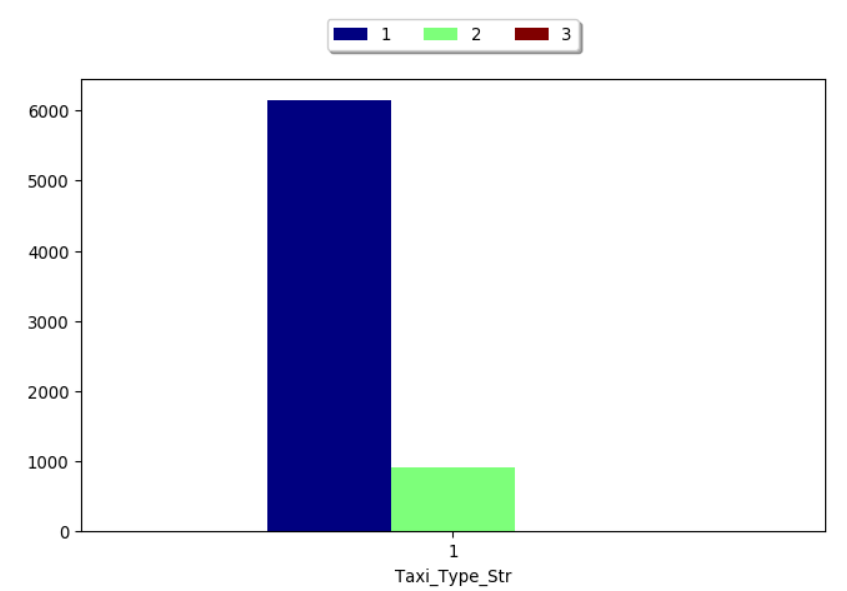
**4.4 Taxi/Private hire car**

****

In the above code I am creating a vTaxi data frame from the join12 data frame. vTaxi contains all the accident details caused by Taxi/Private hire car.

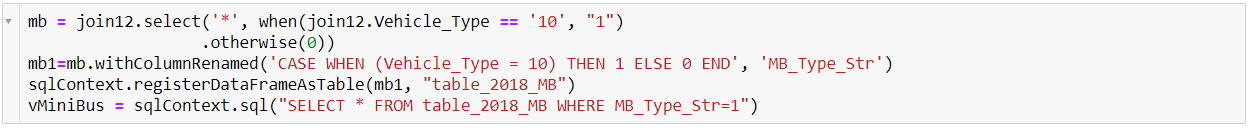


When I display vTaxi the above graph is obtained. The graph represents the accident severity with number of casualties clustered by urban or rural data. For the graph we can say that the total number of casualties is way less compared to the other vehicle types we have compared above. The maximum number just goes above 5000 in the slight cases in the urban region.



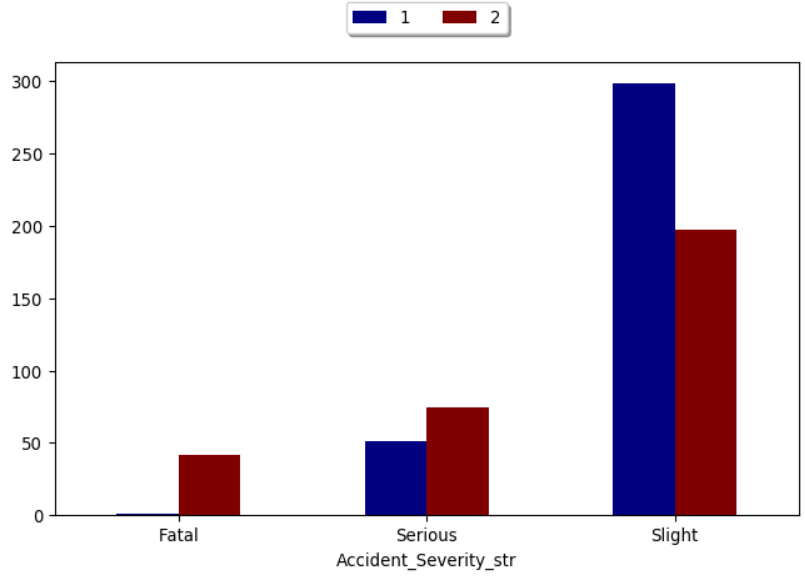
From the above graph we can clearly see that most casualties caused by the Taxi is in the Urban region (Represented by the Blue bar). This is probably because the number of Taxies running in the Rural Region (Represented by the Green bar) is way less compared to the once seen in Urban Region.

**4.5 Minibus (8 - 16 passenger seats)**

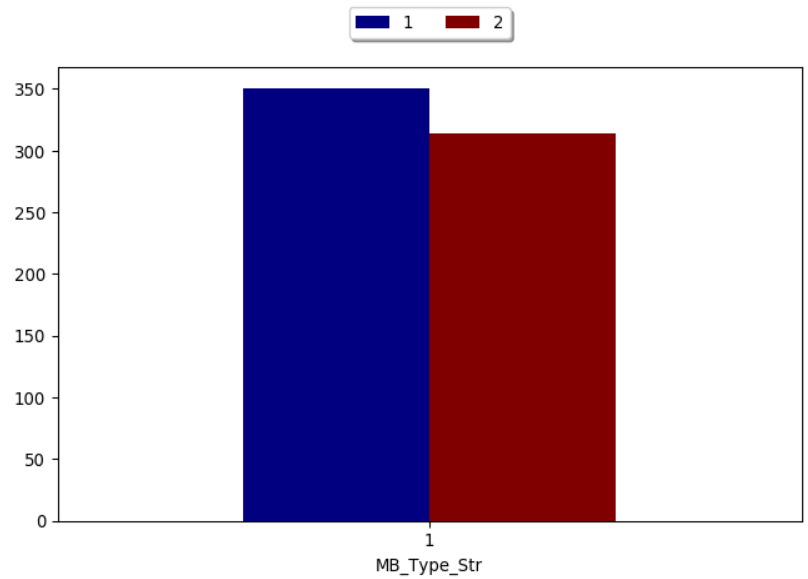


The above code creates a new data frame vMiniBus from join12 which contains all the accidents caused by a Minibus.



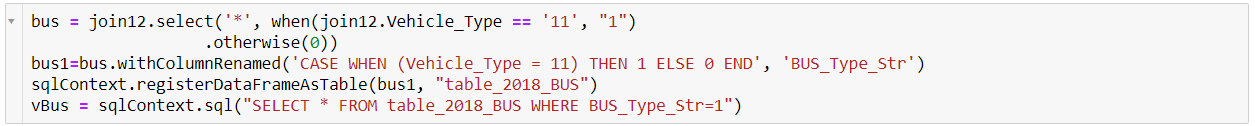


When I display vMiniBus with Accident\_severity and Number of casualties we get the above graph. There are only 405 accidents caused by Minibus which we can see from the casualty number. From the graph we can also see that accidents in the Rural region (represented by the red bar) are more fatal. Accidents caused by Minibus in the Urban region (Represented in Blue) are not fatal as the graph only shows a small amount. This could also be because Minibus driver in the Urban region is more aware compare to the once in the Rural region. The serious accidents are also greater in number in the Rural region compared to the Urban Region. This could indicate that the Minibus drivers in the rural region are not give proper training compared to drivers in the urban region. The slight accidents caused by Minibus is great in the Urban region compared to rural. This could be because of the sheer number of vehicles and the rush found in the Urban region.



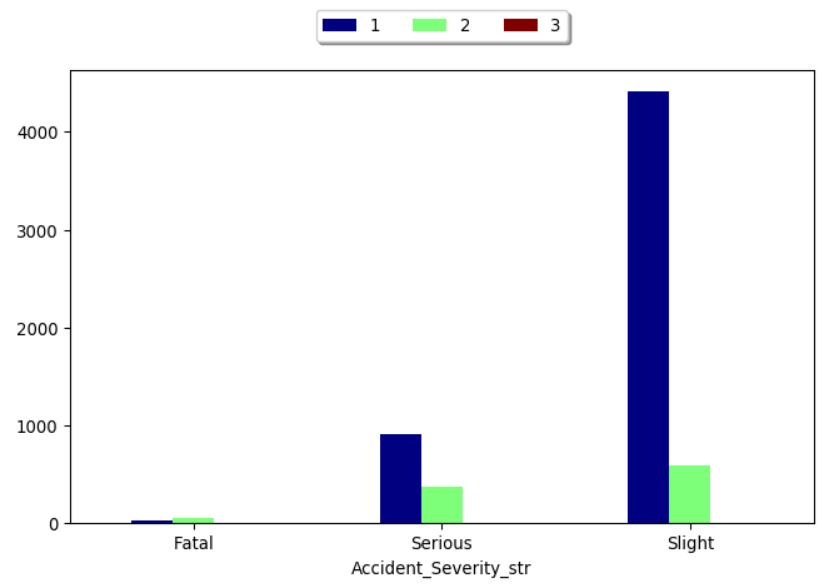
From the above graph we can clearly see that the number of casualties caused by the minibus is more in Urban region (represented in the blue bar) compared to the Rural region (represented in Red bar). But from the graph that shows the Accident\_severty we can clearly tell that fatal and serious accidents occur more in the rural region compared to Urban, but in this graph the data is greater because of the sheer number of slight accidents that occur in the urban region.

**4.6 Bus or coach (17 or more pass seats)**



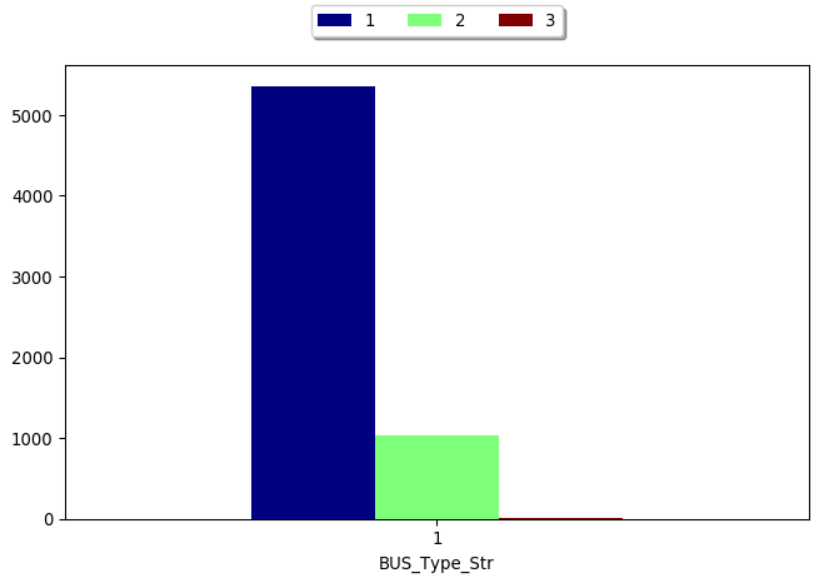
The above code creates a new data frame vBus with the accidents caused by Bus or coach. The total accidents caused by this group is 4513. Now let us take look at the graphs.



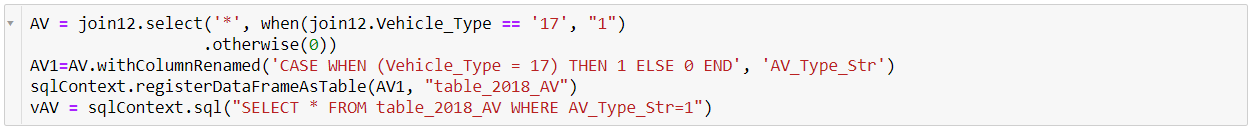


The above graph shows the accident\_severity with the number of casualties caused by the Bus typ

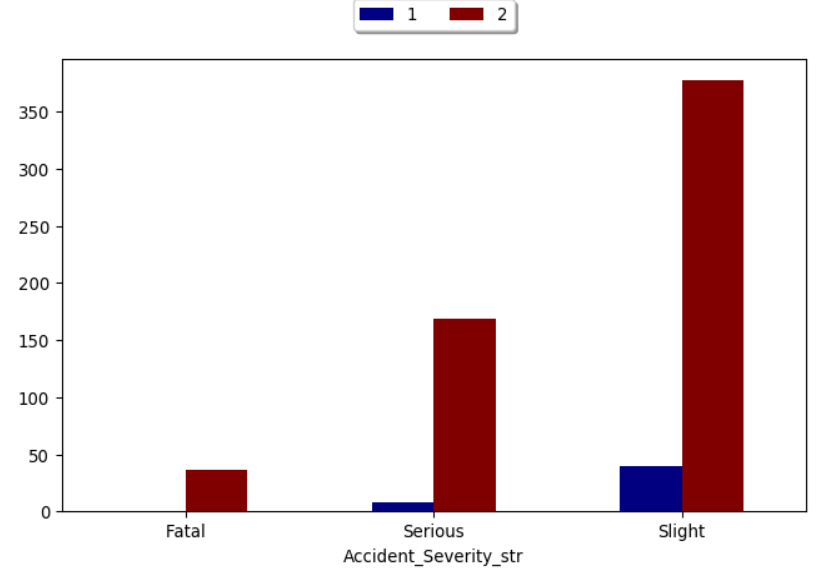
e. We can see that there are more fatal casualties in the rural region (represented by Green bar) compared to the urban region (represented by the blue bar). The number of serious accidents is greater in the urban region compared to rural region. When we look at the Slight accidents, we can see that it is the greatest in the urban region. This could be because it is difficult to operate bus in tight traffic conditions seen in the Urban region and there are chances of slight accidents.



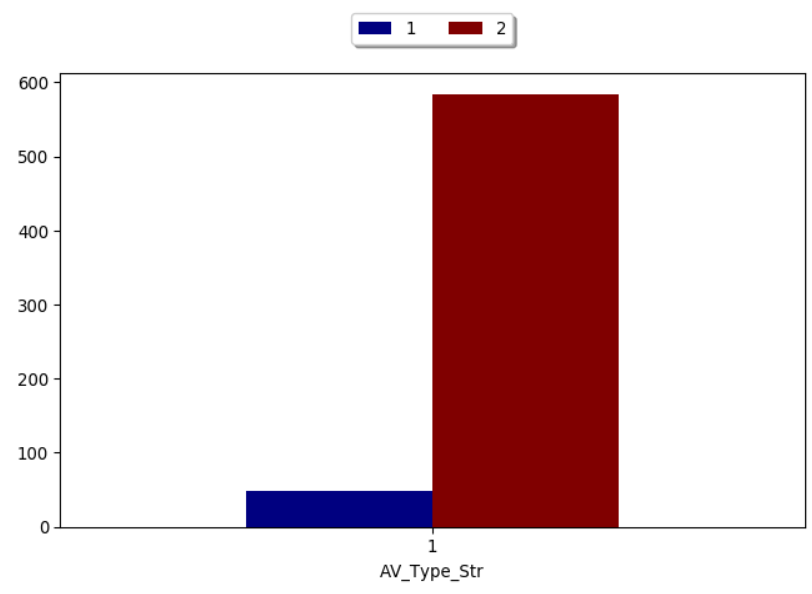
In the above graph I am displaying the BUS\_Type\_Str and Number of casualties. We can see that the Urban region (the blue bar) is has more accidents compared to the rural region (the green bar).

**4.7 Agricultural vehicle**

The above code creates a data frame vAV with the accidents caused my Agricultural vehicles. The data frame is created from the join12 data frame. The total accidents caused by this category is 488

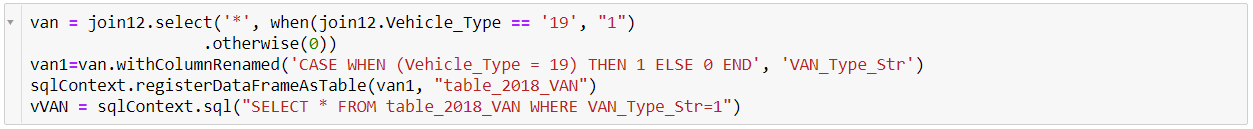


The above graph shows the accident severity and the number of casualties caused by agricultural vehicles. We can see that the major accidents of this vehicle type have occurred in the rural region (represented by the red bar) as agricultural vehicles are mostly used here. We can see that all fatal accidents have occurred in the rural region with nearly 50 fatalities. Some accidents in serious category fall in the urban region (represented by the blue bar). Accidents in the slight category is also seen in the urban region.

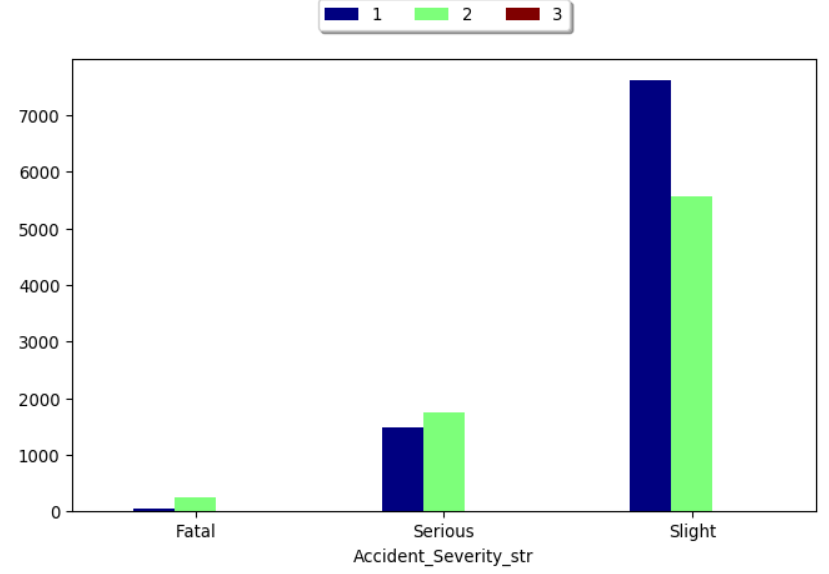


From the graph we can see that the rural region (represented in red bar) has way more casualties compared to the urban region (represented by the blue bar)

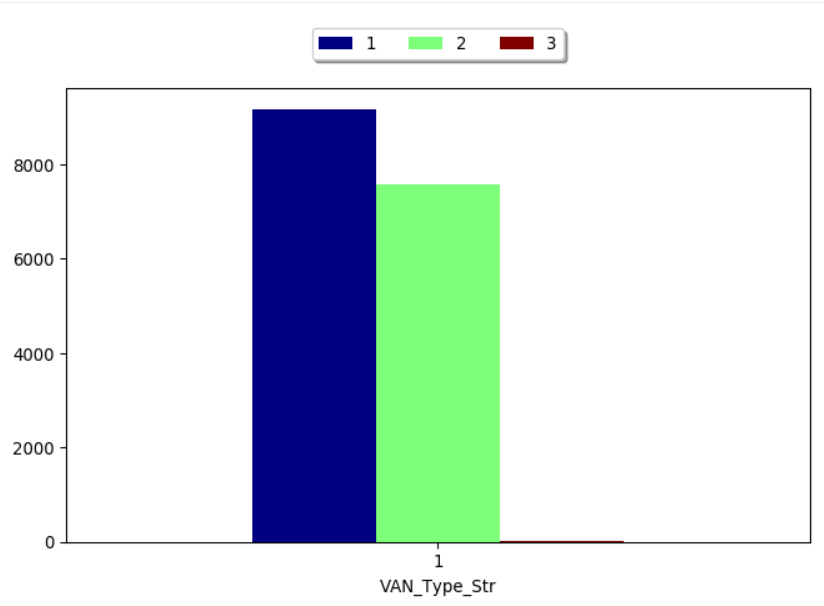
**4.8 Van / Goods 3.5 tonnes mgw or under**



The above code creates a new data frame vVAN which contains the accidents caused by the Van / Goods 3.5 tonnes mgw or under. The total accidents caused by this Vehicle type is 12062.

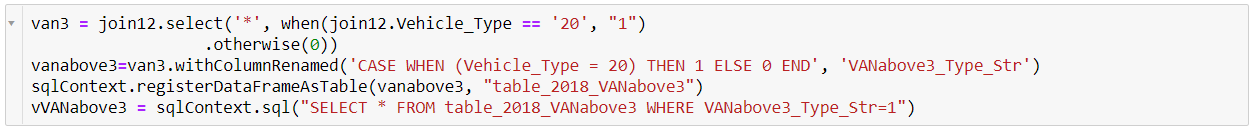


The above graph displays the accident severity and the number of casualties caused by vehicle type Van / Goods 3.5 tonnes mgw or under. We can see that the fatal accidents are greater in the rural region (represented by Green bar) compared to the Urban region (represented by Blue bar). The serious accidents are also greater in the rural region. This could be because this type of vehicle is mainly used in agricultural purposes in the rural region. But the slight accidents occur most in the Urban region. This could be because these types of vehicle are mainly used for transport from rural to urban region.

****

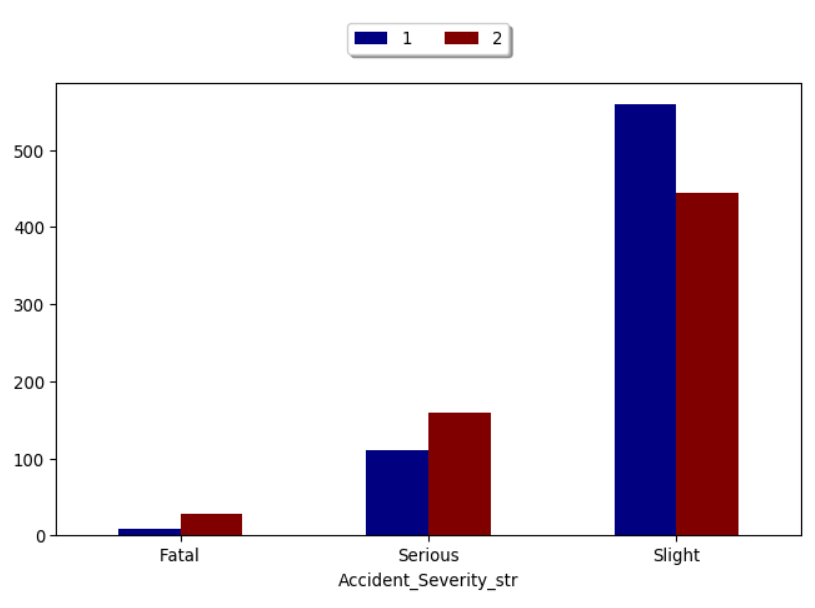
The above graph Shows the casualties caused by VAN type. In this graph we can see that the Urban region (represented in blue bar) has more accidents compared to rural region (represented in Green bar). But from the graph with accident severity, we can see that the greatest number of fatal and serious accidents have occurred in the rural region. In this graph the urban region has a greater value because of the slight accidents that have occurred.

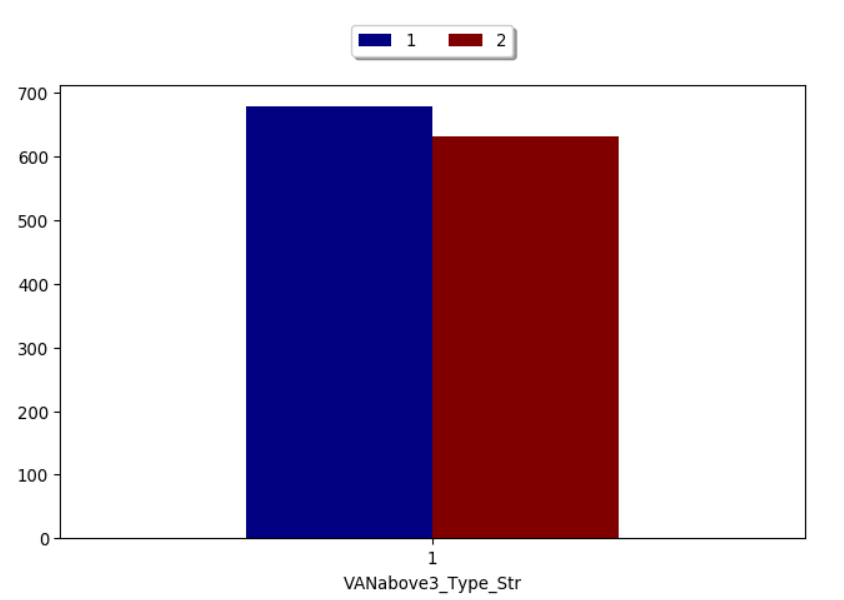
**4.9 Goods over 3.5t. and under 7.5t**



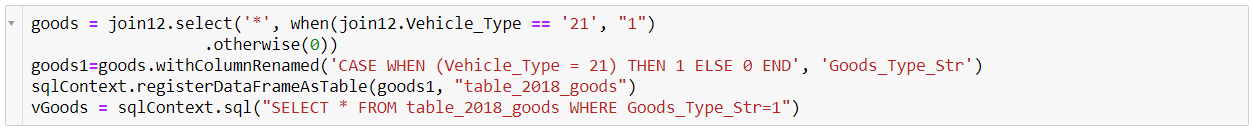
The above code is used to create a new data frame vVANabove3 which has the accidents caused by Goods over 3.5t. and under 7.5t vehicles. The total accidents caused by this type is 1005.



The above graph represents the accident severity and the number of casualties caused by vehicle of the mentioned type. Here we can see that there is a greater fatality number in the rural region (represented by the red bar) compared to the urban region (represented in the blue bar). The serious accidents are also greater in the rural region, but the slight accidents are greater in the urban region.

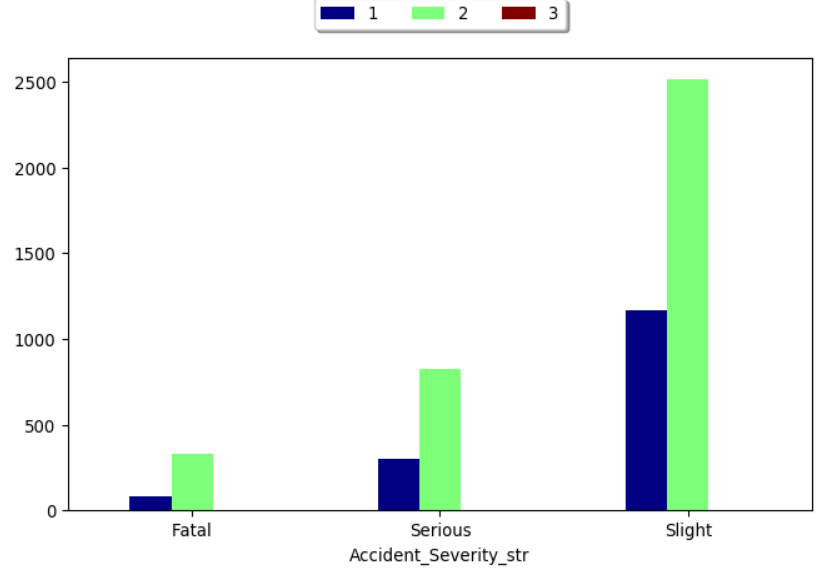


The above graph shows the total casualties caused by Goods over 3.5t. and under 7.5t vehicles in urban (represented by the blue bar) and rural (represented by the red bar) regions. We can see that the urban region has a greater number of casualties compared to the rural region. This could be because of the sheer number of slight accidents in the urban region.

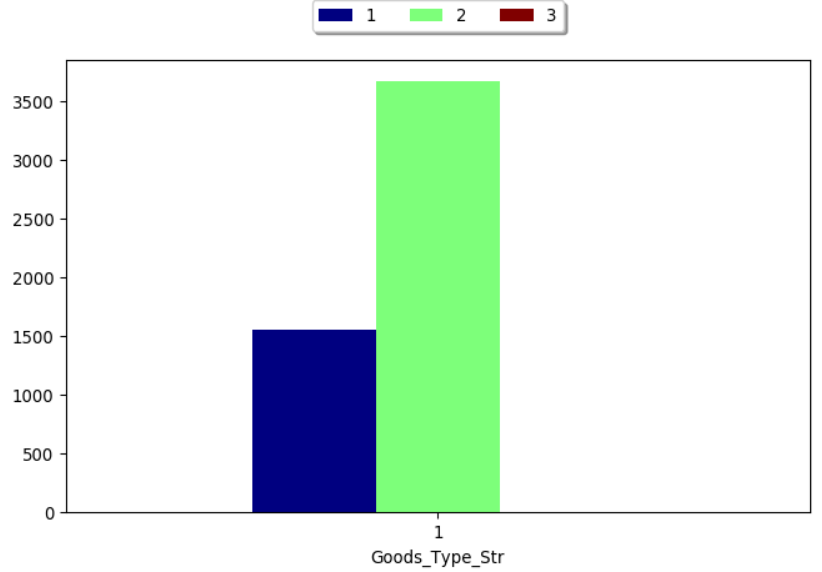
**4.10 Goods 7.5 tonnes mgw and over**

The above code creates a new data frame vGoods that contains the accidents caused by the goods vehicle. The total accidents caused by this type is 3748



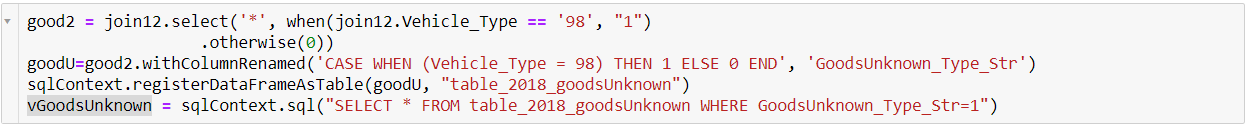


The above graph represents the accident severity and the number of casualties. From the above graph we can see that the rural region (represented in the green bar) has a greater number of accidents in the vehicle type. The fatal serious and the slight accidents are far great in the rural region compared to the urban region (represented in the blue bar). Probably this vehicle is more used in rural regions.



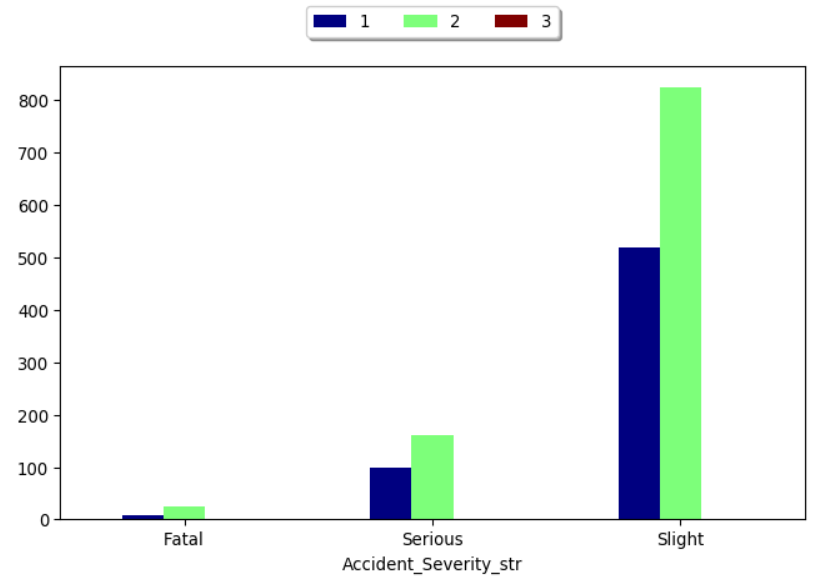
From the above graph we can clearly see how the rural region (represented in the green bar) has more casualties compared to the Urban Region (represented in the blue bar).

**4.11 Goods vehicle - unknown weight**

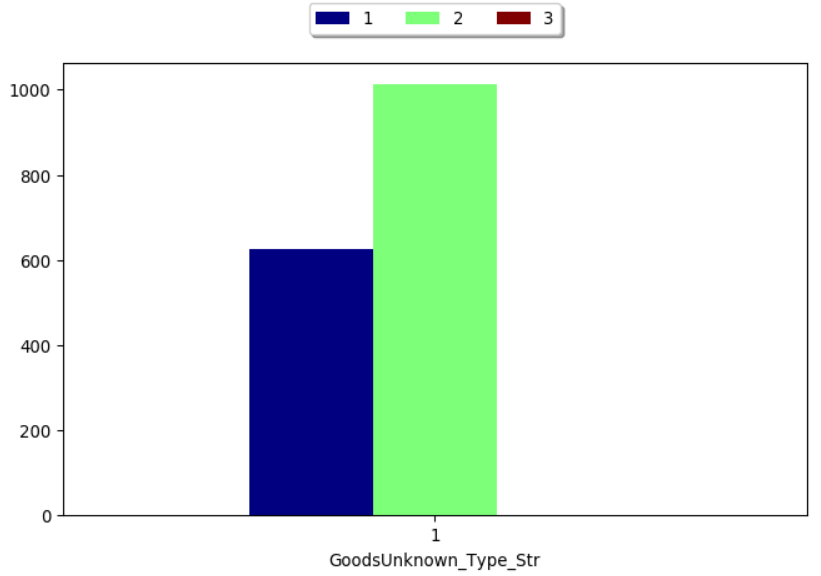
****

The above code creates a new data frame vGoodsUnknown with the accidents caused by Goods vehicle with unknow weight. The total accidents caused by this category is 1205.





The above graph shows the accident severity and number of casualties in the goods unknown vehicle type. From the graph we can see that the number of fatal serious and the slight accidents are the greatest in the rural region (represented by the Green bar) compared to the urban region (represented by the blue bar)



From the above graph we can clearly see that the casualties in the rural region (represented by the Green bar) is greater than the once in the Urban region (represented by the blue bar).

5. CONCLUSION

From the above analysis we can see that both the regions have contributed towards the accidents in 2018, But the highest number of accidents was caused by Cars in the Urban region. Interestingly we can see that accidents caused by pedal bikes comes next and most of them has happened in the Urban region. This indicates a problem in the management of pedal cycles in the urban areas. Most the accidents caused by Goods vehicles are in the Rural region. I was also able to see lots of fatalities in the rural region. This could also be due to the lack of health care facilities in the rural region. From the above analysis I can conclude that the accidents caused in cars are more in the urban region followed by Pedal cycles and Motorcycles while most of the accidents in the rural region are caused by agricultural vehicles and goods vehicles. Therefore, managing these vehicle types and giving proper awareness class for the operators of these vehicles will enhance their awareness and hence reducing the accidents.

6. REFERENCES

https://data.gov.uk/dataset/road-accidents-safety-data

<https://spark.apache.org/docs/latest/sql-ref-syntax-qry-select-case.html>

<https://www.w3schools.com/sql/sql_where.asp>

<https://spark.apache.org/docs/latest/sql-ref-syntax.html>

<https://spark.apache.org/docs/latest/sql-ref-syntax-ddl-create-table-like.html>

<https://spark.apache.org/docs/latest/sql-ref-syntax-qry-select.html>

<https://stackoverflow.com/questions/25157451/spark-sql-case-when-then>

<https://stackoverflow.com/questions/38981405/sql-spark-sql-if-statement-syntax-in-a-query>

<https://pixiedust.github.io/pixiedust/>