**Part B : Pre-recorded Video Presentation (20%)**

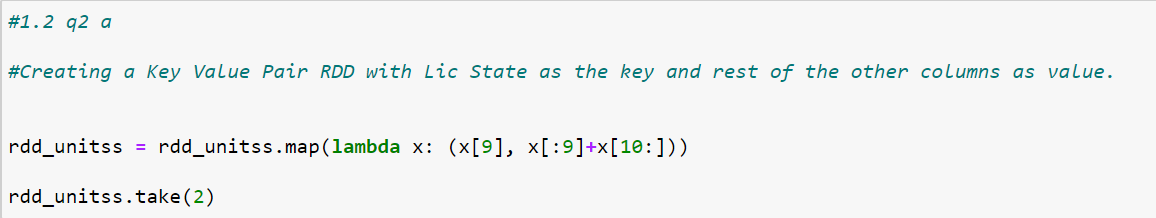
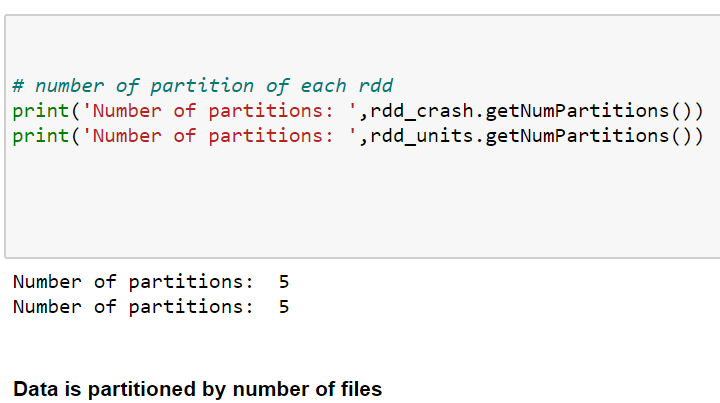
**RDD Partitioning (Task 1.2): Discuss how you implemented the partitioning strategy.**

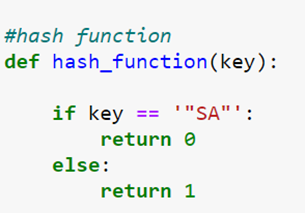
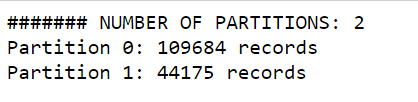
**Also how is the data spread across the partitions after implementing the partitioning**

**strategy? Do you see any skew in data, if so what other approaches can you use to**

**manage the skew?**

After reading the data from the csv files, the data was spread into 5 partitions( one for each of the csv files).

Then I used the map function to split the data by comma and after wards used the map functions again to create a key value pair RDD. I created a hash function that returns 0 is the key is SA, otherwise returns 1. Afterwards I partitioned the RDD into 2 partitions using the hash function. This sends all the records with SA key to partition 0 and the rest to partition 1. Thus, I was able to utilize hash partitioning in this case to answer the given question.



After implementing the hash function, there are 109684 records for “SA” and 44175 records for the rest. We can see that more data is partitioned into SA which means skewness is present. There are many ways to improve skewness while partitioning data:

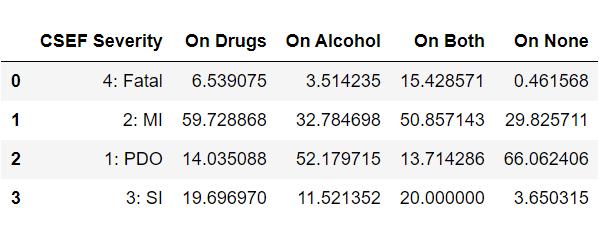
1. Adding more partitions – downside is not helpful when a few keys are dominant
2. Custom partitioning
3. First fit packing
4. Skewed joins
5. Iterative broadcast joins

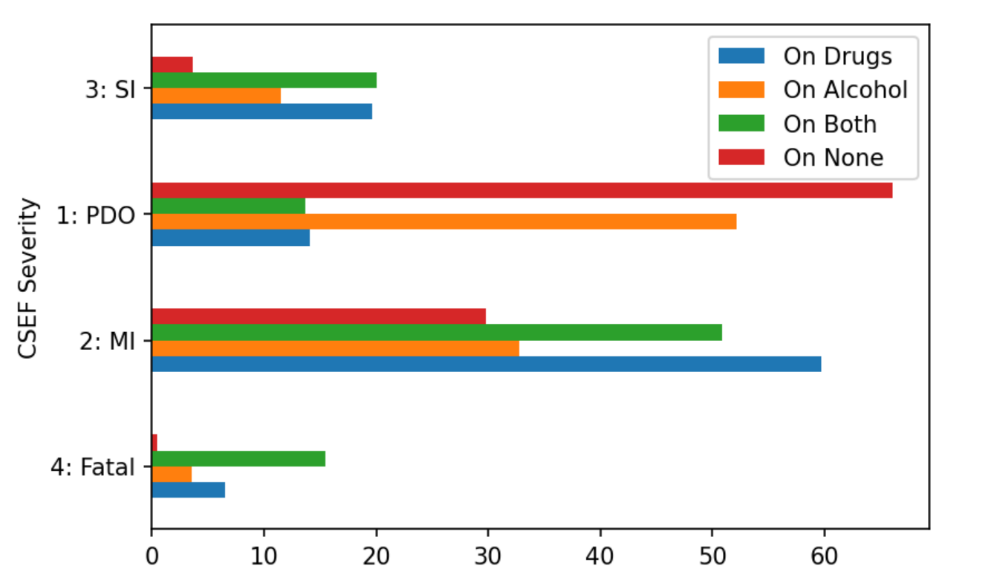
**Crash Severity Analysis (Task 2.3): Based on the results from 2.3.2, combine the**

**results from the four different scenarios to a single table (as shown below) and visualize**

**it with a bar graph.**

**Output table**

****

**Matplotlib grouped bar graph**

The above diagram is a grouped bar chart created from the table using matplotlib (code included in the assignment file). Observations from the graph include:

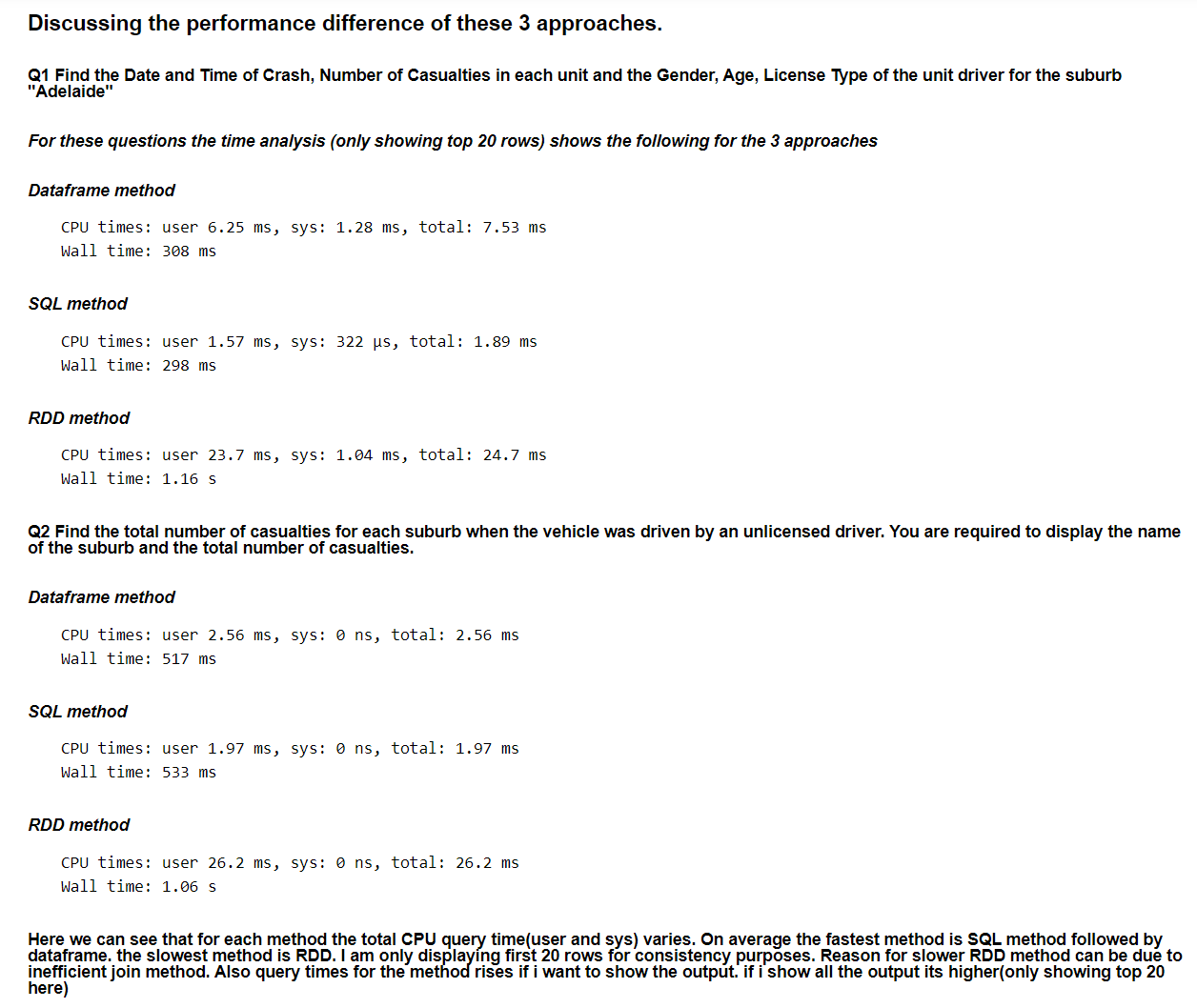
1. PDO severity level has high percentages for people who tested positive for blood alcohol levels. This makes sense as alcohol increases chances of accidents.
2. Minor injury is mostly caused due to drug consumption or people who are consuming both drugs and alcohol. We can see similar trend for severe injury cases.
3. For Fatal injuries, we can observe that the main cause is the consumption of both alcohol and drugs. Only a small number of fatal injuries are caused by nothing. This is due to the negative effects drugs have on the human body

In conclusions we can observe clearly that, drugs and alcohol consumption become the main cause as the severity level of motor vehicle accidents increase.

**RDDs vs DataFrame vs Spark SQL(Task 2.4): Based on the results from 2.4 (1 & 2),**

**present your findings. Explain which approach works faster and the possible reasons**

**for the same.**

****