Project No 2

The New York School authority collects data from all schools that provide bus facilities to students. This data helps to understand if buses are reaching on time or not. This also helps to understand if there is a specific route where buses are taking more time so that it can be improved.  
 You are given a dataset of buses which got broke down or are running late.  
You have been given the below data set:

Filename: bus-breakdown-and-delays.csv

1. School\_Year  
      a. Indicates the year the record refers to  
2. Run\_Type  
     a. Designates whether a breakdown or delay occurred on a specific category of bus services  
3. Bus\_No  
4. Route\_Number  
5. Reason  
     a. Reason for the delay as entered by the staff employed by the reporting bus vendor  
6. Occurred\_On  
7. Number\_Of\_Students\_On\_The\_Bus

You are hired as a big data consultant to provide important insights. You must write a Spark job using the above data and need to provide the following:

1.Load data and create Spark data frame

2.Most common reasons for either a delay or breaking down of the bus  
3. Top five route numbers where the bus was either delayed or broke down  
4. The total number of incidents, year-wise, when the students were Not on the bus  
5. The year in which accidents were less

6.how many incidents are done in 2015-2016 and 2017-2018

7. Year \_wise number of incident