FIT 5202 Assignment 1 Feedback Sheet

Student Name: ISOBEL ROWE

Marked By: David _C.

Part A : Working with RDDs and DataFrames						
Tasks		**Criteria**	**Yes**	**Partial**	**No**	**Comments**
1.1 Data Preparation and Loading	**1**	No of processors and title of application	☑	☐	☐	
		Created SparkSession and SparkConf	☐	☑	☐	- No SparkConf implemented - Additional configuration unnecessary (e.g. mongodb connectors)
	2	Correctly imported the RDDs for Units	☑	☐	☐	
	3	Correctly imported the RDDs for Crashes	☑	☐	☐	
	4	Headers removed, count and first 10 rows displayed for both RDDs	☑	☐	☐	- Rows displayed would be better to show in tabular format
1.2 Data Partitioning in RDD	**1**	Correct number of partitions displayed	☑	☐	☐	
		Default partition strategy answered correctly	☐	☑	☐	- Correct partition strategy, but not explaining the reason of 5 partitions
	2	a. Key Value Pair RDDs created correctly	☑	☐	☐	
		b. Hash function correctly defined	☑	☐	☐	
		b. Partitioning implemented correctly	☑	☐	☐	
		c. Number of records correctly displayed in each partition	☑	☐	☐	
		c. Skewness in partition discussed correctly	☐	☐	☑	- No attempt
1.3 Query/Analysis	**1**	Valid drivers filtered out	☐	☑	☐	- Filtering criteria is not proper (e.g. pedestrian not filtered out), suggest to examine the data to identify the wrong data value
		Average Age calculated correctly	☐	☑	☐	- Incorrect values due to missing filter
	2	Valid vehicle filtered out	☑	☐	☐	
		Oldest and newest vehicles calculated correctly	☑	☐	☐	- There are multiple vehicles with max and min year. Only showed 1 for each of them
2.1 Data Preparation and Loading	**1**	Data loaded into dataframes correctly	☑	☐	☐	

2.1 Data Preparation and Loading	**2**	Schema correctly displayed	☑	☐	☐	
2.2 Query/Analysis	**1**	Filters implemented correctly and data displayed	☑	☐	☐	
	2	10 crash events with highest casualties correctly displayed	☐	☑	☐	- Incorrect results because 'Total Cas' was sorted alphabetically rather than numerically. Column should have been converted to int or float
	3	Total fatalities for each crash type displayed	☑	☐	☐	
		Unlicenced driver filter implemented	☑	☐	☐	
		join, group by and aggregation correctly implemented	☑	☐	☐	
		Results correctly displayed	☑	☐	☐	
2.3 Severity Analysis	**1**	Group by and count correctly implemented	☑	☐	☐	
		Most common severety level answered correctly	☑	☐	☐	
	2	a. Positive on drugs only calculated correctly	☑	☐	☐	
		b. Positive on alc. only calculated correctly	☑	☐	☐	
		c. Positive on both calculated correctly	☑	☐	☐	
		d. Negative for both calculated correctly	☑	☐	☐	
		Brief explanation of the observation	☐	☐	☑	- No attempt in notebook
2.4 RDDs vs DataFrame vs SparkSQL	**1**	Correct implementation for RDD	☐	☐	☑	- No attempt
		Correct implementation for DataFrame	☐	☐	☑	- No attempt
		Correct implementation for Spark SQL	☐	☐	☑	- No attempt
	2	Correct implementation for RDD	☐	☐	☑	- No attempt
		Correct implementation for DataFrame	☑	☐	☐	
		Correct implementation for Spark SQL	☑	☐	☐	
		Discussion of the performance differences	☐	☐	☑	- No attempt in notebook

Qualitative Aspect		Organization of tasks in jupyter notebook Adherance to python standards Use of appropriate comments, output readability	☑	☐	☐	- To increase the readability in the notebook, you should only show the relevant columns, all columns would make it messy and decrease the readability
Part B : Pre-recorded Video Presentation						
1. RDD Partitioning	Partitioning strategy, data distribution, data skewness, approaches to manage skew		☑	☐	☐	- How to manage skewness was not fully explained
2. Crash Severity Analysis	Correctness of the observations based on the bar graph visualization		☑	☐	☐	
3. RDDs vs DataFrame vs Spark SQL	Performance findings , explaination for df/sql being faster than RDD		☐	☑	☐	- DataFrame and SQL queries run faster thanks to optimized execution plans and custom memory management. Suggest doing more reading on those and taking a look at Spark UI to see what happens
Qualitative Aspect	Overall quality of presentation delivery (video/audio clarity) content (use of graphs/quality of slides)		☐	☑	☐	- Suggest adding interaction with the slide, so that it could be easier for the audience to follow your talk - More graphs for slides would be better to be added
		Final Grade	Late Submission		2	**C**