Assignment 1 – Data Analytics with Hive

Introduction

In this project you will be working with the car.csv dataset that you can download from https://www.kaggle.com/mirosval/personal-cars-classifieds

This dataset has the classified records for several Eastern European countries over several years. Beware that the data is not "clean" and investigating and cleaning the data is an important part of the Assignment.

Problem Background

You are the data analyst at a large investment firm that is **contemplating to invest in a used car business**. Your task is to **provide data driven advice** to the stakeholders, that will enable them to **make a sound investment decision**. Failure to make the best decision may result in large financial consequences and irreversible damage to the company reputation and brand.
Your manager has instructed you to use the cars.csv dataset, because the veracity of this data has been established.

Analysis (Questions)

Some analysis questions that you may try to answer

1. What is the relationship between car makes, models and price?
2. What are the top five vehicle manufacturers would you recommend? Why?
3. Does fuel type have any impact on the car price? Explain

Deliverables

You are expected to provide a report of the solutions to the above questions. Each solution must be accompanied by the code used, supporting evidence of its use e.g. relevant output screenshots and detailed method(s) and justification for the given solution. The final report should preferably be presented in a PDF format.

Marks breakdown

30 % - Data cleaning

40 % - Analysis

30 % - Report