POINTS TO NOTE

1. You are required to work on 2 problems – one on regression and the other on classification.
2. Conduct a thorough analysis of the data before you start building the model. Understand the importance of each variable and if you think it should be included in your model or not.
3. Your deliverables would be:
   - R Notebook/Markdown style report with embedded code, plots, comments and sections in it.
   - A presentation which would be evaluated on Saturday, 22nd July in second half of the day.
4. A good report/presentation would be one which has:
   a. Efforts on exploring the data with basic and meaningful plots. (Visualization is not taught to you yet, but use your basic knowledge on what needs to be plotted. The class on visualization on the next day would be much more meaningful then).
   b. Good explanation on how the data is processed and cleaned
   c. Explanation on choice of the model and what are pros/cons of the model you used.
   d. What you would do if you have more time with the problem to improve the model you made.

PROBLEM 1: REGRESSION

You and your team are consulting on an automobile manufacturing company. You are assigned with a task of helping them with the right pricing of the car based on the data they provide. To build an effective model, you are given the following dataset which has several variables representing specs of the car and the price which can be used as target variables. Use linear regression to build a strong model which you can use in future to determine price of the car.

Number of Instances: 205
Number of Attributes: 19
Target Attribute: "price"
Dataset Name: "Automobiles"

Attribute Information

1. normalized-losses: continuous from 65 to 256.
2. aspiration: std, turbo.
3. drive-wheels: 4wd, fwd, rwd.
4. wheel-base: continuous from 86.6 to 120.9.
5. length: continuous from 141.1 to 208.1.
6. width: continuous from 60.3 to 72.3.
7. height: continuous from 47.8 to 59.8.
8. curb-weight: continuous from 1488 to 4066.
9. num-of-cylinders: eight, five, four, six, three, twelve, two.
10. engine-size: continuous from 61 to 326.
11. fuel-system: 1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
12. bore: continuous from 2.54 to 3.94.
13. stroke: continuous from 2.07 to 4.17.
14. compression-ratio: continuous from 7 to 23.
15. horsepower: continuous from 48 to 288.
16. peak-rpm: continuous from 4150 to 6600.
17. city-mpg: continuous from 13 to 49.
18. highway-mpg: continuous from 16 to 54.
19. price: continuous from 5118 to 45400.

Note: Missing Attribute Values: (denoted by "?")

PROBLEM 2: CLASSIFICATION

Here the goal for your team is to predict whether a movie will win an (at least one) academy award or not. You can brain-storm and decide on the target attribute.

People would like to assess the greatness of a movie before it is released in cinema. Therefore, they rely on critics to gauge the quality of a film, while others use their instincts or past experiences. But it takes time to obtain a reasonable number of critic's review after a movie is released & human instinct sometimes is unreliable. Given that thousands of movies are released each year, better ways have been established to assess the greatness of movie. IMDb stores information and other details of the movies and its rating.

Bonus points question: How would your model perform if you were to predict academy award in any one category? (i.e. best actor or best picture or best director)

Use logistic regression to solve the problem.

Number of Instances: 651
Number of Attributes: 32
Dataset Name: "imdb_rotten_tom"

1.title: Title of movie
2. title_type: Type of movie (Documentary, Feature Film, TV Movie)
3. genre: Genre of movie (Action & Adventure, Comedy, Documentary, Drama, Horror, Mystery & Suspense, Other)
4. runtime: Runtime of movie (in minutes)
5. mpaa_rating: MPAA rating of the movie (G, PG, PG-13, R, Unrated)
6. studio: Studio that produced the movie
7. thtr_rel_year: Year the movie is released in theaters
8. thtr_rel_month: Month the movie is released in theaters
9. thtr_rel_day: Day of the month the movie is released in theaters
10. dvd_rel_year: Year the movie is released on DVD
11. dvd_rel_month: Month the movie is released on DVD
12. dvd_rel_day: Day of the month the movie is released on DVD
13. 1imdb_rating: Rating on IMDB
14. imdb_num_votes: Number of votes on IMDB
15. critics_rating: Categorical variable for critics rating on Rotten Tomatoes (Certified Fresh, Fresh, Rotten)
16. critics_score: Critics score on Rotten Tomatoes
17. audience_rating: Categorical variable for audience rating on Rotten Tomatoes (Spilled, Upright)
18. audience_score: Audience score on Rotten Tomatoes
19. best_pic_nom: Whether or not the movie was nominated for a best picture Oscar (no, yes)
20. best_pic_win: Whether or not the movie won a best picture Oscar (no, yes)
21. best_actor_win: Whether or not one of the main actors in the movie ever won an Oscar (no, yes) – note that this is not necessarily whether the actor won an Oscar for their role in the given movie
22. best_actress win: Whether or not one of the main actresses in the movie ever won an Oscar (no, yes) – not that this is not necessarily whether the actresses won an Oscar for their role in the given movie
23. best_dir_win: Whether or not the director of the movie ever won an Oscar (no, yes) – not that this is not necessarily whether the director won an Oscar for the given movie
24. top200_box: Whether or not the movie is in the Top 200 Box Office list on BoxOfficeMojo (no, yes)
25. director: Director of the movie

26. actor1: First main actor/actress in the abridged cast of the movie
27. actor2: Second main actor/actress in the abridged cast of the movie
28. actor3: Third main actor/actress in the abridged cast of the movie
29. actor4: Fourth main actor/actress in the abridged cast of the movie
30. actor5: Fifth main actor/actress in the abridged cast of the movie
imdb_url1 Link to IMDB page for the movie
32. rt_url: Link to Rotten Tomatoes page for the movie