MovieLens_Report

Ian Gledhill

03/06/2019

Introduction

Overview

This report documents the process and methodology adopted by a project to devise a movie recommendation system. The movie recommendations would analyze a dataset containing a large number of reviews of many films by many users.

The users are randomly selected and are only identified by a uniquely assigned user id; there is no demographic and geopgraphic location information used.

Dataset

Source

The source dataset was created by the University of Minnesota and contains:

- 10000054 ratings
- 95580 tags
- 10681 movies by
- 71567 users

The source data has been extracted from the online movie recommendation service MovieLens (https://movielens.org).

The source files can be obtained from this location:

- https://grouplens.org/datasets/movielens/10m/

This link shows further details, including usage rights provided by the license:

- http://files.grouplens.org/datasets/movielens/ml-10m-README.html

Sample Size for Analysis

This is a very large dataset and for the sake of managability a 10% sample of the source dataset will be used by the project.

Shape of the Data

The source data contains three data sets:

- Movies
- Ratings
- Tags

Training and Test Data Partitions

The full approach is described later in this report but in summary the sample data is split into two partitions. A *training partition* will be allocated *90%* of the sample data and will be used to create (or *train*) the algorithm. A *test partition* will be allocated the remaining *10%* of the sample data and will be used to validate the predictions of the algorithm.

Goal

The objective of the project was to find a movie recommendation algorithm. Its success will be measured by reference to the RMSE produced when the algorithm is validated against the test dataset. The RMSE will be graded against a predfined set of thresholds to determine the overall quality of the algorithm (described later).

Approach

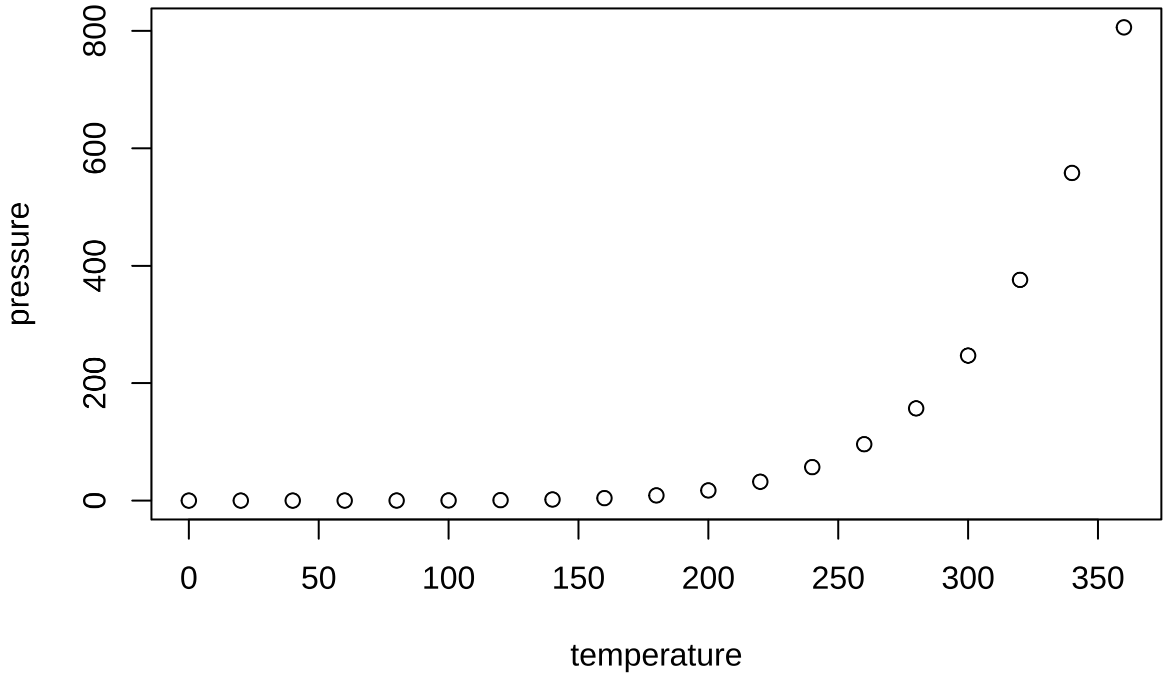
Data Cleansing

Modelling Techniques

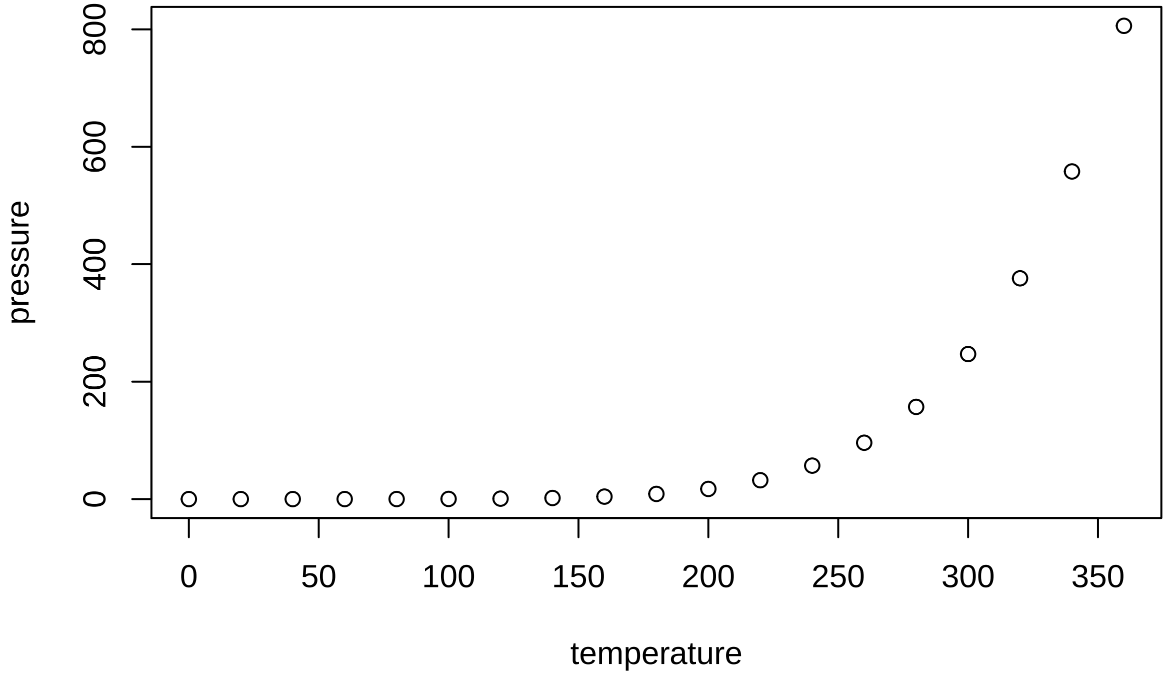
Results

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```



Conclusion


pressure
0
200
400
600
800
0
50
100
150
200
250
300
350
temperature

Bibliography and References