Coursework: EDA & Regression

Mark Muldoon <mark.muldoon@manchester.ac.uk> and
Diego Perez Ruiz <diego.perezruiz@manchester.ac.uk>

28 October – 22 November 2024

This year's coursework involves a synthetic dataset, `MavenRail.csv`, derived from one developed for a visualisation contest run by Maven Analytics, a firm that provides training in Data Science. It describes fictitious rail journeys made by passengers in the UK during the period 1 January–30 April 2024. Some of the submissions to Maven's contest are available at

`https://bit.ly/MavenRailChallenge`

You needn't look at them as part of the assignment, but the variety of visualisations on display is interesting and impressive.

Table 1 shows some typical records and leads to certain immediate observations.

- Many of the variables here are *categorical* rather than numeric. That is, they take on a modest number of distinct values that have no natural ordering. Examples include `Railcard`, `Departure.Station` and `Journey.Status`. When doing EDA about such variables, it is often more informative to make tables rather than, say, to draw bar charts. The pandas function `crosstab()` and the built-in R function `xtabs()` make it easy to prepare such tables.
- Ticket prices are given in pounds.
- Three of the columns, `Departure`, `Scheduled.Arrival` and `Actual.Arrival`, involve dates and times in the format

  `DD/MM/YYYY HH:mm`

  where `D`, `M`, `Y`, `H` and `m` represent, respectively, digits specifying day, month, year, hour and minute. Both Python (via Pandas) and R have built-in capabilities to handle this representation of time: see the jupyter notebook `HandlingDates.ipynb` or the R script `HandlingDates.R` for examples. Both are available in the same place as this assignment.

Row Number	1	2	3	4
Payment.Method	Contactless	Credit Card	Credit Card	Credit Card
Railcard	Adult	Adult	None	None
Ticket.Class	Standard	Standard	Standard	Standard
Ticket.Type	Advance	Advance	Advance	Advance
Price	43	23	3	13
Departure.Station	London Paddington	London Kings Cross	Liverpool Lime Street	London Paddington
Arrival.Station	Liverpool Lime Street	York	Manchester Piccadilly	Reading
Departure.Datetime	2024-01-01 11:00	2024-01-01 09:45	2024-01-02 18:15	2024-01-01 21:30
Scheduled.Arrival.Datetime	2024-01-01 13:30	2024-01-01 11:35	2024-01-02 18:45	2024-01-01 22:30
Actual.Arrival.Datetime	2024-01-01 13:30	2024-01-01 11:40	2024-01-02 18:45	2024-01-01 22:30
Journey.Status	On Time	Delayed	On Time	On Time
Reason.for.Delay		Signal Failure		
Refund.Request	No	No	No	No

Table 1: *Illustrative data from Maven's catalog of rail journeys. The table has been transposed (rows and columns swapped) so that it will fit on the page.*

Prepare a 1000 word report that summarises your work on the following exercises.

1. Write a brief description of the data, including its origin. You should imagine you are writing for a group who have no idea what this dataset is about. *[1 mark]*
2. Do an exploratory data analysis. *[5 marks]*
3. Add a column, `DelayInMinutes`, to the dataset that gives the duration of the delay in minutes, if the journey was delayed. If the train arrived on time, set `DelayInMinutes` to NA. *[2 marks]*
4. Restrict attention to those journeys where `Journey.Status` is *not* `On Time` and add a column, `MediumPrice`, to the resulting dataset that answers the question: Does the ticket price lie in the range
$$£10 < \texttt{Price} \leq £30?$$
Fit an appropriate regression model that predicts whether a passenger will request a refund using `MediumPrice` as a single predictor. *With the help of the fitted model*, answer the following questions (show your calculations, either by hand or with help of R or Python):
   - What is the probability that a passenger will request a refund, given that they paid £5 for their ticket?
   - What is the probability that a passenger will request a refund, given that their ticket cost £25?

   [4 marks]
5. Using the data in `MavenRail.csv`, fit appropriate regression models and use them to determine how likely the passengers whose data appear in the file `ToPredict.csv` are to request a refund. You are free to choose which explanatory variables to include in your model and may, if you like, compare several models, but make sure that you state clearly your final choice of model and give reasons supporting this choice. With the help of your chosen model, interpret the results in terms of probability of requesting a refund (as you did for the model based on `MediumPrice`). *[6 marks]*
6. Include (in an appendix) all R or Python code used to produce the analysis. *[2 marks]*

Illustrate your analysis with appropriate figures and tables. Figure and table captions, the contents of tables and your code do not count against the word limit.

Due Date: 17:00 on 22 November 2024, uploaded to BlackBoard as a PDF. Also note:

- We want to mark your work anonymously, so please don't include your name in your report. Instead, label it with your student ID number.
- Although there is no minimum or maximum number of references required, you should reference any sources (except for materials from this course) that you use when developing your code or preparing your report. The list of references should come at the end of the report and does not count against the word limit.

Late work, word limits, plagiarism and all that:
Penalties for lateness and for reports that are too long or too short, as well as the university's rules about how to acknowledge your sources, are discussed in:

- The University-wide Policy on Submission of Work for Summative Assessment on Taught Programmes.
- The University-wide Guidance to students on plagiarism and other forms of academic malpractice. This includes a discussion of how to acknowledge the use of Generative Artificial intelligences such as ChatGPT (we discourage you from using such tools: learn to write your own text and code) and links to further resources, including an article Can I use a chatbot or AI tool in my assignments? in the Library's FAQ.
- The School of Social Sciences Handbook.