Applied Statistics
Programming Assignment

Read the instructions carefully before you submit your report. Reports that are not typeset (e.g., photographs of handwritten pages) or are not in pdf will NOT be graded. You can form teams of up to three people. All members of the group must submit the same file to be graded. The file name of your report shoud be lastname1_lastname2_lastname3.pdf. For example, if my team includes only myself, my report will be named triantafyllou.pdf.

In this project, we will try to predict the box office revenue (in dollars) of movies. The data include 3000 movies in The movie database. The data include information like the movie's budget, language, genre, and cast. Some of are numerical random variables (e.g., budget), some are categorical (e.g., genre) and some information is in text format. Your goal is to predict the worldwide gross revenue (variable name: revenue) using the information in the data. A detailed description of the data can be found in the last page of this document.
The data are available in the course website. You can implement the assignment in any programming language you want. Your main deliverable is a typewritten (not a photograph of a handwritten) report **in pdf format**, that includes your answers to questions below. The answers can be supported by plots and/or results copied from your analysis. Your report should have the following components: An introductory paragraph, describing the goals of your analysis, and one section per exercise. Each section must summarize your findings and conclusions for the corresponding exercise.
You must also include your code, but I will not necessarily run it. Do not copy code, because I will run it through code similarity detection software. You can submit your document and accompanying code on the course website, by Sunday, June 18th, 23:59. Late submissions will not be accepted.

1. (20 points) We will first consider only the following explanatory variables:

   - budget
   - binary variable denoting if the movie is english or not.
   - running time
   - popularity

   For each of the numerical explanatory variables, compute the correlation coefficient and visualize the relationship of the variable with the response variable with a scatter plot. If you could only use one variable to predict revenue, which one would you use?

2. (80 points) You will now use multiple regression to predict movie revenue. You can use just the variables in Exercise 1 above, or you can construct additional features: For example, you could construct a binary variable encoding if Brad Pitt is in the cast, or a numerical variable that encodes the number of female actors in the film.
   (a) Briefly describe your analysis (no more than 2 paragraphs). Did you include additional variables, and, if so, which ones? What is the R^2 of your model?
   (b) Are the conditions for the multiple linear regression model satisfied? Explain your answer.
   (c) What is the slope associated with the variable you identified as most predictive in Exercise 1 above? What is the interpretation of the slope?
   (d) Which variables are significant for predicting movie revenue?

Διαβάστε προσεκτικά τις οδηγίες πριν υποβάλετε την αναφορά σας. Αναφορές που δεν είναι δακτυλογραφημένες (π.χ. φωτογραφίες χειρόγραφων σελίδων) ή δεν είναι σε pdf ΔΕΝ θα βαθμολογηθούν. Μπορείτε να σχηματίσετε ομάδες έως τριών ατόμων. Όλα τα μέλη της ομάδας πρέπει να υποβάλουν το ίδιο αρχείο για να βαθμολογηθούν. Το όνομα αρχείου της αναφοράς σας θα πρέπει να είναι lastname1 lastname2 lastname3.pdf. Για παράδειγμα, εάν Η ομάδα μου περιλαμβάνει μόνο εμένα, η αναφορά μου θα ονομαστεί triantafyllou.pdf.

Σε αυτό το έργο, θα προσπαθήσουμε να προβλέψουμε τα έσοδα από τα box office (σε δολάρια) των ταινιών. Τα δεδομένα περιλαμβάνουν 3000 ταινίες στη βάση δεδομένων The movie database. Τα δεδομένα περιλαμβάνουν πληροφορίες όπως ο προϋπολογισμός της ταινίας, η γλώσσα, το είδος και το καστ. Μερικές από αυτές είναι αριθμητικές τυχαίες μεταβλητές (π.χ. προϋπολογισμός), κάποιες είναι κατηγορικές (π.χ. είδος) και κάποιες οι πληροφορίες είναι σε μορφή κειμένου. Ο στόχος σας είναι να προβλέψετε τα παγκόσμια ακαθάριστα έσοδα (όνομα μεταβλητής: revenue) χρησιμοποιώντας τις πληροφορίες στα δεδομένα. Λεπτομερή περιγραφή των δεδομένων μπορείτε να βρείτε στην τελευταία σελίδα.
Τα δεδομένα είναι διαθέσιμα στον ιστότοπο του μαθήματος. Μπορείτε να υλοποιήσετε την εργασία σε οποια γλώσσα προγραμματισμού θέλετε. Το κύριο παραδοτέο σας είναι μια δακτυλόγραφη (όχι φωτογραφία χειρόγραφης) αναφοράς μορφή pdf, που περιλαμβάνει τις απαντήσεις σας στις παρακάτω ερωτήσεις. Μπορείτε να συμπεριλάβετε γραφήματα ή και αποτελέσματα αντιγραμένα από την ανάλυσή σας. Η αναφορά σας πρέπει να έχει τα ακόλουθα στοιχεία: Εισαγωγική παράγραφο, που περιγράφει τους στόχους της ανάλυσής σας, και μία ενότητα ανά άσκηση. Κάθε ενότητα πρέπει να συνοψίζει τα ευρήματά σας και τα συμπεράσματά σας για την αντίστοιχη άσκηση. Πρέπει επίσης να συμπεριλάβετε τον κωδικό σας, αλλά δεν θα τον τρέξω απαραίτητα. Μην αντιγράψετε κώδικα, γιατί θα τον ελέγξω μέσω λογισμικού ανίχνευσης ομοιότητας κώδικα. Μπορείτε να υποβάλετε το έγγραφό σας και τον συνοδευτικό κωδικό σας στην ιστοσελίδα του μαθήματος, έως την Κυριακή 18 Ιουνίου, 23:59. Εκπρόθεσμες υποβολές δεν θα γίνονται δεκτές.

1. (20 points) Αρχικά θα εξετάσουμε μόνο τις ακόλουθες επεξηγηματικές μεταβλητές:

   - προϋπολογισμός (budget)
   - δυαδική μεταβλητή που υποδηλώνει εάν η ταινία είναι αγγλική ή όχι.
   - διάρκεια (running time)
   - δημοτικότητα (popularity)

   Για καθεμία από τις αριθμητικές επεξηγηματικές μεταβλητές, υπολογίστε τον συντελεστή συσχέτισης και απεικονίστε τον σχέση της μεταβλητής με τη μεταβλητή απόκρισης με ένα scatter plot. Εάν μπορούσατε να χρησιμοποιήσετε μόνο μία μεταβλητή για να προβλέψετε έσοδα, ποιο θα χρησιμοποιούσατε;

2. (80 points) Τώρα θα χρησιμοποιήσετε πολλαπλή γραμμική παλινδρόμηση για να προβλέψετε τα έσοδα μιας. Μπορείτε να χρησιμοποιήσετε μόνο τις μεταβλητές στην Άσκηση 1 παραπάνω, ή μπορείτε να δημιουργήσετε πρόσθετες μεταβλητές: Για παράδειγμα, θα μπορούσατε να δημιουργήσετε μία δυαδική μεταβλητή που να κωδικοποιει το αν ο Brad Pitt είναι στο καστ, ή μια αριθμητική μεταβλητή που μετράει το ποσοστό των γυναικών ηθοποιών στο καστ.
   (a) Περιγράψτε εν συντομία την ανάλυσή σας (όχι περισσότερες από 2 παραγράφους). Έχετε συμπεριλάβει επιπλέον μεταβλητές, και, αν ναι, ποιες; Ποιο είναι το R^2 του μοντέλου σας;
   (b) Ικανοποιούνται οι προϋποθέσεις για το μοντέλο πολλαπλής γραμμικής παλινδρόμησης; Εξήγησε την απάντησή σου.
   (c) Ποια είναι η κλίση που σχετίζεται με τη μεταβλητή που προσδιορίσατε ως πιο προγνωστική στην Άσκηση 1 παραπάνω; Ποια είναι η ερμηνεία της κλίσης;
   (d) Ποιες μεταβλητές είναι σημαντικές για την πρόβλεψη των εσόδων μίας ταινίας;

- Data Description id - Integer unique id of each movie.
- *belongs_to_collection* - Contains the TMDB Id, Name, Movie Poster and Backdrop URL of a movie in JSON format.
- *budget* -Budget of a movie in dollars. 0 values mean unknown.
- *genres*- Contains all the Genres Name & TMDB Id in JSON Format
- *homepage* - Contains the official homepage URL of a movie. Example: http://sonyclassics.com/whiplash/, this is the homepage of Whiplash movie.
- *imdb_id* - IMDB id of a movie (string). You can visit the IMDB Page like this: https://www.imdb.com/title/
- *original_language* - Two digit code of the original language, in which the movie was made. Like: en = English, fr = french.
- *original_title* - The original title of a movie. Title & Original title may differ, if the original title is not in English.
- *overview* - Brief description of the movie.
- *popularity* - Popularity of the movie in float.
- *poster_path* - Poster path of a movie. You can see the full image like this: https://image.tmdb.org/t/p/original/.
- *production_companies* - All production company name and TMDB id in JSON format of a movie.
- *production_countries* - Two digit code and full name of the production company in JSON format.
- *release_date* - Release date of a movie in mm/dd/yy format.
- *runtime* - Total runtime of a movie in minutes (Integer).
- *spoken_languages* - Two digit code and full name of the spoken language.
- *status* - Is the movie released or rumored?
- *tagline* - Tagline of a movie
- *title* - English title of a movie
- *Keywords* - TMDB Id and name of all the keywords in JSON format.
- *cast* - All cast TMDB id, name, character name, gender (1 = Female, 2 = Male) in JSON format
- *crew* - Name, TMDB id, profile path of various kind of crew members job like Director, Writer, Art, Sound etc.
- *revenue* - Total revenue earned by a movie in dollars.