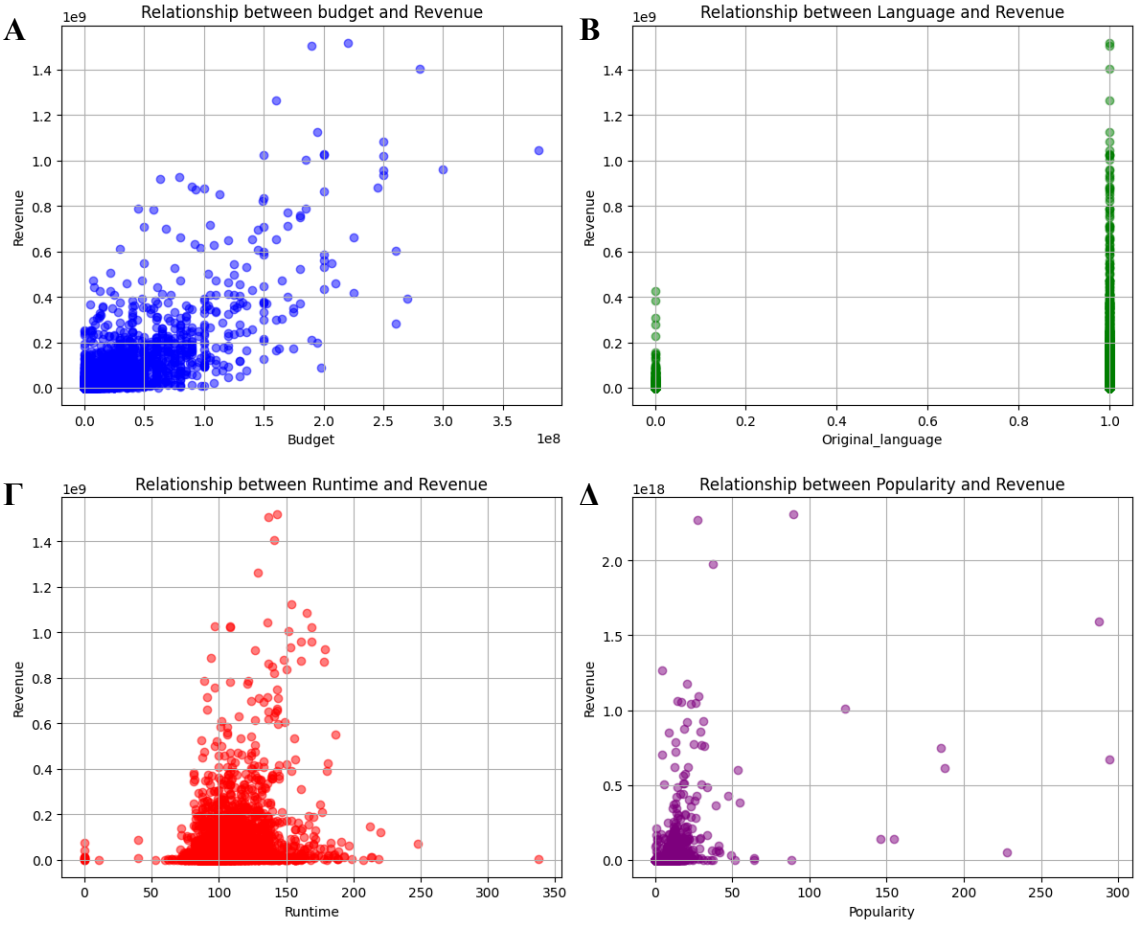
Εισαγωγή

Η βιομηχανία κινηματογράφου είναι ένας τομέας που αποφέρει δισεκατομμύρια στην παγκόσμια οικονομία και βασίζεται κυρίως στην πρόβλεψη των κατάλληλων συνθηκών έτσι ώστε μία ταινία να αποφέρει το μεγαλύτερο δυνατό κέρδος. Η επιτυχία μιας ταινίας μπορεί να εξαρτηθεί από διάφορους παράγοντες όπως το είδος την ταινίας, τον προϋπολογισμό, τη γλώσσα και τους ηθοποιούς. Αυτή η αναφορά, έχει ως στόχο να χρησιμοποιήσει όλα αυτά τα δεδομένα για να κατασκευάσει ένα μοντέλο το οποίο θα μπορεί να προβλέψει τα παγκόσμια ακαθάριστα έσοδα (revenue) μία ταινίας. Η προσοχή μας θα αφορά κυρίως την κατανόηση της σχέσης ανάμεσα στα χαρακτηριστικά που καθορίζουν μία ταινία και τη μεταβλητή απόκρισης (revenue). Τα δεδομένα τα οποία θα χρησιμοποιηθούν για να διεξαχθεί η ανάλυση, απαρτίζονται από ένα αρχείο που περιέχει πληροφορίες για 3000 ταινίες.

Ερώτημα 1

Στο ερώτημα αυτό, μας απασχολούν τέσσερις μόνο επεξηγηματικές μεταβλητές από το αρχείο: ο προϋπολογισμός, η δυαδική μεταβλητή που υποδηλώνει εάν η ταινία είναι αγγλική ή όχι, η διάρκεια και η δημοτικότητα. Στο αρχείο μας, η μεταβλητή που υποδηλώνει εάν η ταινία είναι αγγλική ή όχι δεν ήταν σε δυαδική μορφή, οπότε πριν την ανάλυση χρειάστηκε μετατροπή σε δυαδική.

Από κάτω φαίνονται τα scatter plots που απεικονίζουν τη σχέση κάθε μεταβλητής με τη μεταβλητή απόκρισης:



(**Α**): Παρατηρούμε μία θετική συσχέτιση ανάμεσα στον προϋπολογισμό και τα παγκόσμια ακαθάριστα έσοδα καθώς όσο αυξάνεται ο προϋπολογισμός τείνουν επίσης να αυξάνονται και τα έσοδα. Η συσχέτιση όμως δεν είναι απόλυτα γραμμική καθώς υπάρχουν ταινίες οι οποίες παρόλο τον μεγάλο προϋπολογισμό παρουσιάζουν μικρά έσοδα.
(**Β**): Παρατηρούμε ότι οι ταινίες με κύρια γλώσσα τα αγγλικά (δυαδική απεικόνιση 1.0) τείνουν να έχουν περισσότερα έσοδα από τις ταινίες με κάποια διαφορετική κύρια γλώσσα (δυαδική απεικόνιση 0.0).
(**Γ**): Παρατηρούμε μία μικρή θετική συσχέτιση μεταξύ διάρκειας και εσόδων. Οι ταινίες με διάρκεια από 75 έως και περίπου 150 λεπτά παρουσιάζουν μεγαλύτερα έσοδα από πολύ μικρής/μεγάλης διάρκειας ταινίες.
(**Δ**): Παρατηρούμε ότι το γράφημα είναι αρκετά διασπαρμένο καθώς ταινίες με μεγάλη αλλά και μικρή δημοτικότητα παρουσιάζουν υψηλά αλλά και χαμηλά έσοδα.

Ο συντελεστής συσχέτισης κάθε επεξηγηματικής μεταβλητής από τις παραπάνω με τα παγκόσμια ακαθάριστα έσοδα μιας ταινίας είναι ο εξής:

corr(budget,revenue) = 0.7529645103815288
corr(original_language,revenue) = 0.1421298728540003
corr(runtime,revenue) = 0.21638013018147206
corr(popularity,revenue) = 0.4614602896736129

Μπορούμε να παρατηρήσουμε ότι τη μεγαλύτερη συσχέτιση με τα παγκόσμια ακαθάριστα έσοδα παρουσιάζει ο προϋπολογισμός μιας ταινίας. Σε συνδυασμό με τα scatter plots που ο προϋπολογισμός είναι ο μόνος που παρουσιάζει ισχυρή θετική συσχέτιση με τα παγκόσμια ακαθάριστα έσοδα μπορούμε να συμπεράνουμε ότι ο προϋπολογισμός είναι η καλύτερη μεταβλητή για να κάνουμε πρόβλεψη για τα έσοδα.

Ερώτημα 2

(α)

Στην ανάλυσή μου χρησιμοποιώ πολλαπλή γραμμική παλινδρόμηση για να προβλέψω τα έσοδα μίας ταινίας βασιζόμενη στις επεξηγηματικές μεταβλητές του προηγούμενου ερωτήματος. Υπολόγισα το R τετράγωνο με τις επεξηγηματικές μεταβλητές του προηγούμενου ερωτήματος και ήταν 0,615. Στην προσπάθεια μου να βελτιώσω αυτή την τιμή, έψαξα στο αρχείο με όλες τις ταινίες που μελετάμε, τον ηθοποιό ο οποίος έπαιξε στις περισσότερες από αυτές και βρήκα τον Samuel L. Jackson. Έτσι, αν πρόσθετα άλλη μία δυαδική μεταβλητή στο μοντέλο για το αν έπαιζε ο συγκεκριμένος ηθοποιός στην ταινία και έκανα ξανά την πρόβλεψη το R στο τετράγωνο έπαιρνε την τιμή 0.621 που είναι σαφώς καλύτερη. Στη συνέχεια έψαξα να βρω τον δεύτερο ηθοποιό που έπαιξε στις περισσότερες ταινίες, τον Robert De Niro, ο οποίος όμως σε συνδυασμό με τον Samuel L. Jackson έφερναν τιμή μικρότερη του 0.621. Τέλος, από προσωπικό θαυμασμό

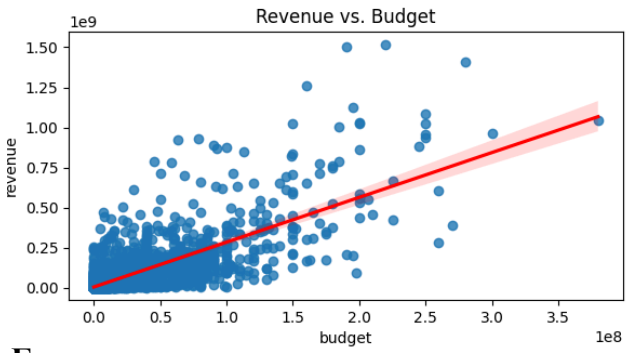
προς την ηθοποιό Anne Hathaway δοκίμασα να βάλω στο μοντέλο την δυαδική μεταβλητή για το αν έπαιζε στην ταινία ο Samuel L. Jackson ή η Anne Hathaway και πήρα την μεγαλύτερη τιμή για το R τετράγωνο ίση με 0,622.

(β)

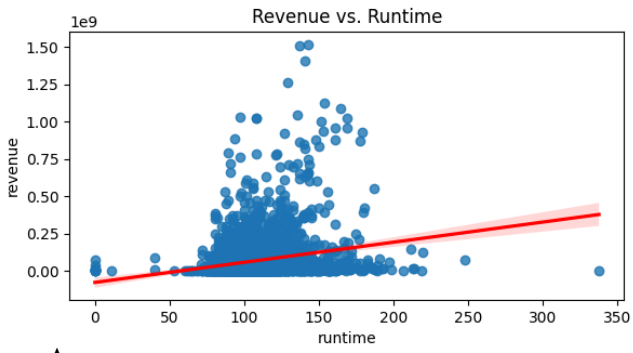
Οι προϋποθέσεις που πρέπει να ικανοποιούνται για το μοντέλο πολλαπλής γραμμικής παλινδρόμησης είναι οι εξής: κάθε επεξηγηματική μεταβλητή πρέπει να σχετίζεται γραμμικά με τα έσοδα, τα residuals πρέπει να παρουσιάζουν κανονική κατανομή, να είναι ανεξάρτητα μεταξύ τους και να έχουν σταθερή μεταβλητότητα.

Για να ελέγξουμε τη γραμμικότητα φτιάχνουμε γραφήματα που συσχετίζουν κάθε μεταβλητή με τη μεταβλητή απόκρισης:

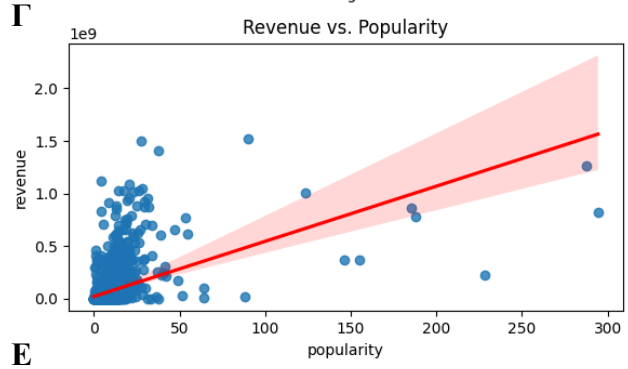
A



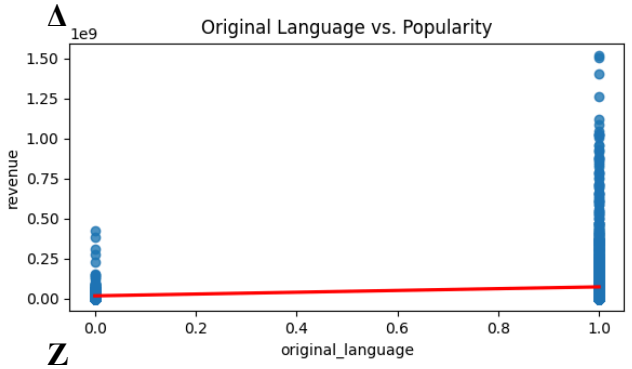
B



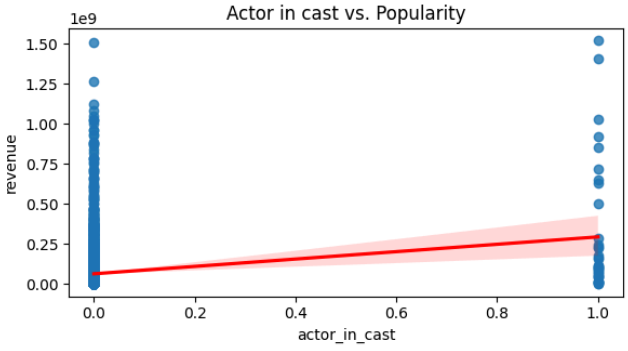
Γ



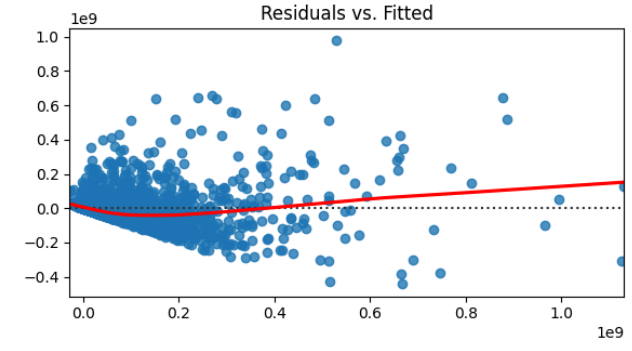
Δ



E



Z

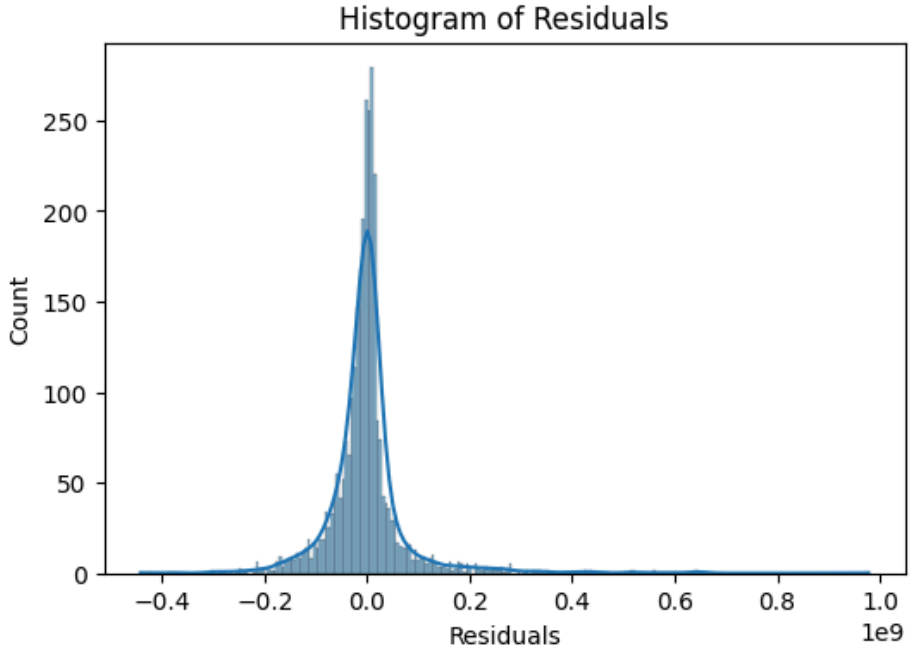


(A-E): Παρατηρούμε ότι κάθε επεξηγηματική μεταβλητή που χρησιμοποιούμε στο μοντέλο μας παρουσιάζει γραμμική σχέση με τη μεταβλητή απόκρισης, ειδικότερα ο προϋπολογισμός και η δημοτικότητα.
(Z): Σύμφωνα με το γράφημα, τα υπολείμματα δεν παρουσιάζουν σταθερή μεταβλητότητα καθώς υπάρχει μία ανομοιόμορφη διασπορά των residuals στις προσαρμοσμένες τιμές.

Η τιμή Durbin-Watson για το κατασκευασμένο μοντέλο είναι ίση με περίπου 2,017 το οποίο δείχνει ότι δεν υπάρχει σημαντική συσχέτιση μεταξύ των residuals και άρα είναι ανεξάρτητα.

Παράλληλα, στο παρακάτω ιστόγραμμα παρόλο που το κέντρο της κατανομής βρίσκεται στο μηδέν, τα residuals δεν ακολουθούν κανονική κατανομή καθώς η δεξιά ουρά είναι μακρύτερη από την αριστερή και στο μηδέν παρατηρούνται πολύ μεγαλύτερες τιμές από τις υπόλοιπες. Αυτά τα αποτελέσματα υποστηρίζονται και από τις τιμές των Omnibus and Jarque-Bera τεστ τα οποία αποδεικνύουν ότι η κατανομή διαφέρει από την κανονική.

Επομένως, δεν ικανοποιούνται όλες οι προϋποθέσεις για το μοντέλο πολλαπλής γραμμικής παλινδρόμησης.



(γ)

Στην Άσκηση 1, η μεταβλητή που προσδιορίστηκε ως η πιο προγνωστική για την πρόβλεψη των εσόδων ήταν ο προϋπολογισμός. Η κλίση της μεταβλητής αυτής είναι ίση με τον συντελεστή συσχέτισης της, ο οποίος είναι ίσος με 2,4478. Η τιμή αυτή προσδιορίζει πως εάν όλες οι άλλες επεξηγηματικές μεταβλητές ήταν σταθερές μία αλλαγή κατά μία μονάδα στον προυπολογισμό, αυξάνει κατά περίπου 2,45 μονάδες τα παγκόσμια ακαθάριστα έσοδα (revenue)

(δ)

Για να ελέγξουμε ποιες επεξηγηματικές μεταβλητές του μοντέλου είναι σημαντικές για την πρόβλεψη των παγκόσμιων ακαθάριστων εσόδων θα χρησιμοποιήσουμε την τιμή p, σύμφωνα με την οποία αν μία μεταβλητή έχει τιμή $p < 0,05$ τότε είναι σημαντική για την πρόβλεψη.

Οι μεταβλητές προυπολογισμός (budget), δημοτικότητα (popularity) και ηθοποιός στο καστ (actor_in_cast) έχουν τιμή p ίση με 0,00 άρα είναι πολύ σημαντικές για την πρόβλεψη. Η μεταβλητή για τη διάρκεια της ταινίας έχει p-value ίσο με 0,033 το οποίο είναι μικρότερο από το 0,05 άρα είναι σημαντική. Τέλος, η μεταβλητή για τη γλώσσα της ταινίας έχει τιμή p ίση με 0,337 η οποία είναι μεγαλύτερη από το 0,05 και άρα δεν είναι σημαντική.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.494e+07	9.16e+06	-2.722	0.007	-4.29e+07	-6.98e+06
budget	2.4478	0.046	52.658	0.000	2.357	2.539
original_language	-4.4e+06	4.58e+06	-0.960	0.337	-1.34e+07	4.59e+06
runtime	1.561e+05	7.3e+04	2.137	0.033	1.29e+04	2.99e+05
popularity	2.61e+06	1.36e+05	19.129	0.000	2.34e+06	2.88e+06
actor_in_cast	1.107e+08	1.41e+07	7.841	0.000	8.3e+07	1.38e+08