**Υποερώτημα 4**

Στη συνέχεια θα χρησιμοποιήσουμε την τελειοποιημένη Βάση δεδομένων για την παραγωγή γνώσης. Ο αλγόριθμος που θα χρησιμοποιηθεί είναι η Πολλαπλή Γραμμική Παλινδρόμηση (Multiple Linear Regression) ξεχωριστά για το καθε προϊόν. Θα εφαρμόσουμε την τεχνική στη γλώσσα προγραμματισμού Python χρησιμοποιώντας τα πακέτα: Matplotlib,Numpy,Pandas και Sklearn.

**Λιγα λόγια για τη τεχνική**: Η Πολλαπλή Γραμμική Παλινδρόμηση είναι μια τεχνική στην οποία λαμβάνουμε υπόψη μας μια σειρά από χαρακτηριστικά (features) και μια εξαρτημένη μεταβλητή (target) για να παράξουμε μια γραμμή που εκφράζει τη γενική κίνηση της εξαρτημένης μεταβλητής εν συναρτήσει των ανεξάρτητων. Ο ορισμός της Γραμμής Πολλαπλής Γραμμικής Παλινδρόμησης P για χαρακτηριστικα *x*1, *x*2, ... , *x*p  ειναι:  y = 0 + 1x1 + 2x2 + ... + pxp . Το τελικό αποτέλεσμα εκφράζει το πως η μέση τιμή μ αλλάζει εν συναρτήσει των χαρακτηριστικών. Έτσι, μπορόυμε να μοντελοποιήσουμε την εμπορική συμπεριφορά του προϊόντος και να προβλέψουμε πόσες πωλήσεις θα έχει κατα μέσο όρο στην ίδια ημερομηνία.

**Χαρακτηριστικα και Στόχος (Features and Target)**: Όπως προαναφέραμε τα χαρακτηριστικα της Βάσης μας ειναι τα εξης: Ημέρα (Day), Μήνας (Month) και οι πωλήσεις του καθε προϊοντος ξεχωριστά, τα οποία ευελπιστούμε και να προβλέψουμε. Για τον σκοπό αυτόν, θα πρέπει να εφαρμόσουμε την τεχνική ξεχωριστά για καθε προϊόν καθως το κάθε προϊον έχει διαφορετικές τάσεις πωλήσεων ανάλογα με την ημέρα και τον μήνα. Στη συνέχεια, αφού η εφαρμογή ειναι η ίδια σε κάθε προϊόν, θα εξετάσουμε την εφαρμογή της τεχνικής για το προϊόν “Americano”.

**Εφαρμογή:**

Θα ξεκινήσουμε εισάγωντας τις απαραίτητες βιβλιοθήκες για την λειτουργία της εφαρμογής μας:

import matplotlib.pyplot as plt

import numpy as np

from sklearn import linear\_model, metrics

import csv

import pandas as pd

Μετά θα πρέπει να εισάγουμε την βάση δεδομένων την οποία θα επεξεργαστούμε:

data = pd.read\_csv('Bakery Sum Sales.csv')

Πριν εφαρμόσουμε την τεχνική, πρέπει να διαχωρίσουμε τα δεδομένα που θα χρησιμοποιήσουμε όπως αναφέραμε πιο πριν. Δηλαδή σε Χαρακτηριστικά και Στόχο (που σε αυτή την περίπτωση είναι οι πωλήσεις του “Americano”).

# defining feature matrix(X) and response vector(y)

X = data[['Day','Month']]

y = data['americano']

Για να μπορέσουμε να κρίνουμε το πόσο επιτύχημένο ήταν το εγχείρημά μας θα υιοθετήσουμε μια τεχνική που λέγεται Train/Test Split. Θα διαχωρήσουμε δηλαδή την Βάση Δεδομένων μας σε δυο κομμάτια. Ένα κομμάτι (το 60% των εισαγωγών) για την εκπαίδευση του μοντέλου στο οποίο η Μεταβλητή Στόχου (Target Variable) είναι ορατή και ένα κομμάτι (το 40% των εισαγωγών) για την δοκιμή του μοντέλου στο οποίο η Μεταβλητή Στόχου δεν είναι ορατή. Με αυτόν τον τρόπο επιτυγχάνουμε μια πρακτική δοκιμασία του μοντέλου μας ώστε να ξέρουμε πόσο καλά μπορεί να λειτούργησει σε πραγματικές καταστάσεις.

# splitting X and y into training and testing sets

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.4, random\_state=1)

Καλούμε το μοντέλο μας και εισάγουμε τα δεδομένα εκπαίδευσης.

# create linear regression object

reg = linear\_model.LinearRegression()

# train the model using the training sets

reg.fit(X\_train, y\_train)

Και για να τελειωποιήσουμε το μοντέλο μας, κάνουμε μια σειρά απο μετρήσεις.

# regression coefficients

print('Coefficients: ', reg.coef\_)

# variance score: 1 means perfect prediction

print('Variance score: {}'.format(reg.score(X\_test, y\_test)))

r\_sq = reg.score(X\_train, y\_train)

print(f"coefficient of determination: {r\_sq}")

**Αποτελέσματα:**

***Coefficients: [-0.00072963 0.32523976]***

***Variance score: -0.20838884822708348***

***coefficient of determination: 0.06232276367254075***

**Παράδειγμα πρόβλεψης:**

