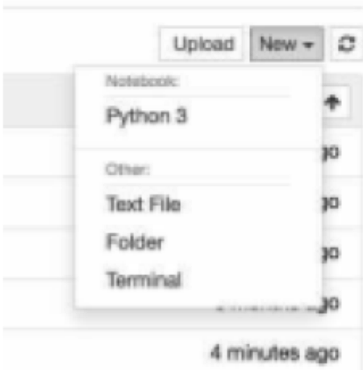
Εισαγωγή αρχείου δεδομένων CMS και δημιουργία ιστογραμμάτων για την ανάλυση δεδομένων

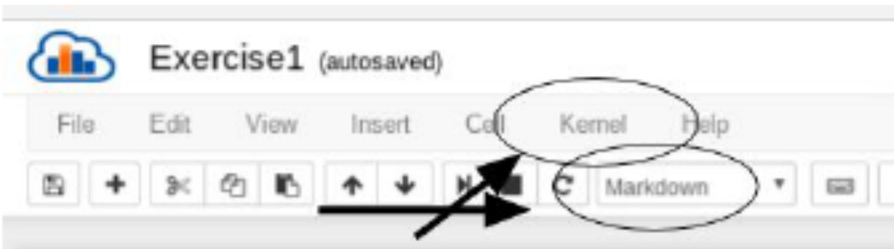
Αυτή η άσκηση χρησιμοποιεί το Jupyter Notebook, το οποίο μπορείτε να κατεβάσετε δωρεάν εγκαθιστώντας το Anaconda (https://www.continuum.io/downloads) για την ανάλυση συνόλων δεδομένων CMS από το CERN.
Ανοίξτε την πύλη δεδομένων, http://opendata.cern.ch/

Επιλέξτε ένα αρχείο δεδομένων για ανάλυση από το http://opendata.cern.ch/record/545. Για παράδειγμα, κατεβάστε το Dimuon_DoubleMu.cvs. Σημειώστε ότι υπάρχουν άλλα αρχεία διαθέσιμα, αλλά το αρχείο πρέπει να έχει κατάληξη (.csv). Κάντε λήψη του επιλεγμένου αρχείου, σημειώνοντας την ακριβή τοποθεσία και το όνομα του αρχείου.

Ανοίξτε το Anaconda - Navigator. Ξεκινήστε το Notebook Jupyter. Επιλέξτε "New" και μετά "Python 3".



Κουτιά που δίνουν οδηγίες ή εξηγήσεις αλλά όχι κωδικοποίηση, μπορεί να εισαχθεί επιλέγοντας "Markdown" από το αναπτυσσόμενο μενού.



Αν θέλετε ναεπιστρέψετε σε μια σελίδα αφού προηγουμένως την έχετε επεξεργαστεί, συνιστάται να επιλέξετε "Kernel" που ακολουθείται από "Restart & Clear Output". Για να εισαγάγετε κώδικα, επιλέξτε "Code" από το αναπτυσσόμενο μενού. Ο κώδικας μπορεί να αντιγραφεί από άλλες πηγές, να επικολληθεί και να επεξεργαστεί.

Πρώτα, θα πρέπει να εισαγάγετε τα πακέτα pandas και matplotlib.pyplot.

```
In [5]: import pandas
        import matplotlib.pyplot as plt
        %matplotlib inline
```

Αφού εισαγάγετε ένα πλαίσιο κώδικα, πατήστε ctrl & enter ταυτόχρονα για να εκτελέσετε τον κώδικα. Ένας αστερίσκος εμφανίζεται στο In [] κατά την επεξεργασία της εντολής. Περιμένετε έως ότου εμφανιστεί ένας αριθμός στην τοποθεσία αυτή πριν προχωρήσετε. Τα μηνύματα σφάλματος θα εμφανιστούν με ροζ χρώμα και θα υποδείξουν πληροφορίες σχετικά με το σφάλμα.

Για να χρησιμοποιήσετε το αρχείο που έχετε κατεβάσει, πρέπει να το αποθηκεύσετε σε μια μεταβλητή. Πληκτρολογήστε τη θέση του αρχείου και το όνομα ακριβώς όπως εμφανίζεται στον υπολογιστή σας. Σε αυτό το παράδειγμα, το αρχείο είναι Data/Dimuon_DoubleMu.cvs.

Αποθηκεύστε τα δεδομένα στο μεταβλητό σύνολο δεδομένων και ελέγξτε τα περιεχόμενα των πρώτων 5 σειρών.

```
In [6]: dataset = pandas.read_csv('data/Dimuon_DoubleMu.csv')

        dataset.head()
```

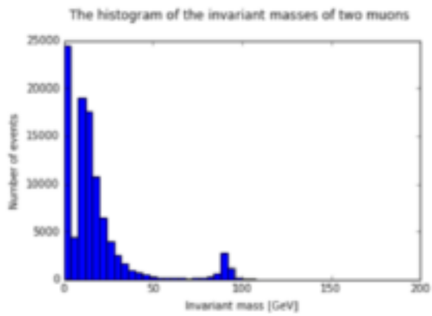
Αν η εντολή έχει γραφτεί σωστά, θα δημιουργηθεί πίνακας δεδομένων. Για το δεδομένο παράδειγμα, η μάζα αδρανείας παρουσιάζει ιδιαίτερο ενδιαφέρον. Για να δημιουργήσετε ένα ιστόγραμμα της μάζας αδράνειας, αποθηκεύστε το σύνολο δεδομένων βάσει της επικεφαλίδας της στήλης ακριβώς όπως εμφανίζεται στο (M) ως μεταβλητή invariant_mass.
Σχεδιάστε ένα ιστόγραμμα, αναφέροντας τη μεταβλητή στο σχέδιο, ορίζοντας τον αριθμό των στοιχείων και το εύρος. Στο παρακάτω παράδειγμα, χρησιμοποιούμε 50 στοιχεία και σχεδιάζουμε από 0 έως 200 GeV.

```
In [7]: invariant_mass = dataset['M']

        plt.hist(invariant_mass, bins=50, range=(0,200))
        plt.show()
```

Αυτό το γράφημα ήταν κυρίως για να ελέγξει τα δεδομένα μας και να σχεδιάσει τις εντολές. Ο αριθμός των στοιχείων και εύρος μπορεί να μεταβληθεί προκειμένου να αναλυθεούν σαφέστερα τα δεδομένα. Παρακάτω υπάρχουν εντολές για να ανατυπώσετε το γράφημα με ετικέτες τίτλου και άξονα.

```
In [6]: plt.xlabel('Invariant mass [GeV]')
        plt.ylabel('Number of events')
        plt.title('The histogram of the invariant masses of two muons \n')

        plt.hist(invariant_mass, bins=50, range=(0,200))
        plt.show()
```

Ένα δείγμα ιστόγραμμα εμφανίζεται πιο κάτω. Το γράφημα δείχνει τα παρακάτω γεγονότα κάτω από50 GeV και με μέγιστη τιμή περίπου 90 GeV.



Μαθηματικές επεμβάσεις, όπως προσθήκη ή αφαίρεση, μπορούν να πραγματοποιηθούν στα δεδομένα ορίζοντας νέες μεταβλητές που επιτρέπουν την περαιτέρω ταξινόμηση των δεδομένων. Στο παρακάτω παράδειγμα, τα αρχικά δεδομένα χωρίζονται σε δύο νέα σύνολα δεδομένων με βάση την ενέργεια τη σύγκρουσης. Σε κάθε ομάδα δεδομένων δίνεται ένα όνομα και οργανώνεται στην περίπτωση αυτή με υψηλή ενέργεια (> 150GeV) και χαμηλής ενέργειας (<150 GeV).

```
In [7]: newsethighE = dataset[dataset.E1+dataset.E2>150]
        newsetlowE = dataset[dataset.E1+dataset.E2<150]
```

Το νέο σύνολο δεδομένων μπορεί να γραφτεί ξεχωριστά, όπως έγινε παραπάνω ή σε μία γραφική παράσταση. Τα δύο ιστογράμματα μπορούν να επικαλυφθούν προσαρμόζοντας τη διαφάνεια χρησιμοποιώντας την εντολή alpha. Οι ετικέτες για κάθε σύνολο δεδομένων περιλαμβάνονται στη λεζάντα που βρίσκεται στην επάνω δεξιά γωνία. Μπορούμε επίσης να αλλάξουμε την εμβέλεια για να επικεντρωθούμε στο γεγονός που μας ενδιαφέρει.

```
In [43]: plt.xlabel('Invariant mass [GeV]')
         plt.ylabel('Number of events')
         plt.title('The invariant masses of two muons comparing high and low energy\n')
         plt.hist(newsetlowE ['M'], bins=50, range=(80,100),alpha=0.5, label='Low E')
         plt.hist(newsethighE ['M'], bins=50, range=(80,100),alpha=0.5, label='High E')
         plt.legend (loc='upper right')
         plt.show()
```

Το τελικό αποτέλεσμα παρουσιάζεται πιο κάτω