Homework 1

Frequent Itemsets from Gene Dataset

L1 - {gene\_1}: sup = 0.83, {gene\_12}: sup = 0.54, {gene\_14}: sup = 0.52, {gene\_17}: sup = 0.55, {gene\_21}: sup = 0.62, {gene\_22}: sup = 0.55, {gene\_23}: sup = 0.54, {gene\_25}: sup = 0.57,

{gene\_26}: sup = 0.52, {gene\_27}: sup = 0.51, {gene\_3}: sup = 0.71, {gene\_31}: sup = 0.51,

{gene\_36}: sup = 0.61, {gene\_37}: sup = 0.56, {gene\_39}: sup = 0.51, {gene\_4}: sup = 0.5,

{gene\_43}: sup = 0.5, {gene\_45}: sup = 0.58, {gene\_47}: sup = 0.66, {gene\_48}: sup = 0.57

{gene\_5}: sup = 0.73, {gene\_50}: sup = 0.5, {gene\_53}: sup = 0.5, {gene\_54}: sup = 0.67,

{gene\_55}: sup = 0.55, {gene\_56}: sup = 0.51, {gene\_59}: sup = 0.76, {gene\_6}: sup = 0.66,

{gene\_60}: sup = 0.54, {gene\_63}: sup = 0.5, {gene\_64}: sup = 0.5, {gene\_66}: sup = 0.59,

{gene\_67}: sup = 0.62, {gene\_71}: sup = 0.58, {gene\_72}: sup = 0.74, {gene\_75}: sup = 0.57

{gene\_77}: sup = 0.58, {gene\_78}: sup = 0.59, {gene\_8}: sup = 0.66, {gene\_81}: sup = 0.58

{gene\_83}: sup = 0.5, {gene\_84}: sup = 0.54, {gene\_87}: sup = 0.67, {gene\_89}: sup = 0.59,

{gene\_9}: sup = 0.5, {gene\_90}: sup = 0.52, {gene\_91}: sup = 0.65, {gene\_93}: sup = 0.53

{gene\_94}: sup = 0.62, {gene\_98}: sup = 0.51, {gene\_99}: sup = 0.56

L2 - {gene\_21, gene\_1}: sup = 0.53, {gene\_1, gene\_3}: sup = 0.63, {gene\_1, gene\_47}: sup = 0.59

{gene\_5, gene\_1}: sup = 0.65, {gene\_54, gene\_1}: sup = 0.58, {gene\_59, gene\_1}: sup = 0.62

{gene\_1, gene\_6}: sup = 0.59, {gene\_67, gene\_1}: sup = 0.55, {gene\_1, gene\_72}: sup = 0.61

{gene\_8, gene\_1}: sup = 0.53, {gene\_81, gene\_1}: sup = 0.51, {gene\_84, gene\_1}: sup = 0.5

{gene\_87, gene\_1}: sup = 0.56, {gene\_1, gene\_89}: sup = 0.52, {gene\_91, gene\_1}: sup = 0.55

{gene\_94, gene\_1}: sup = 0.54, {gene\_47, gene\_3}: sup = 0.5, {gene\_5, gene\_3}: sup = 0.59

{gene\_59, gene\_3}: sup = 0.56, {gene\_72, gene\_3}: sup = 0.53, {gene\_5, gene\_47}: sup = 0.53

{gene\_5, gene\_59}: sup = 0.51, {gene\_5, gene\_6}: sup = 0.52, {gene\_5, gene\_72}: sup = 0.51

{gene\_5, gene\_87}: sup = 0.51, {gene\_5, gene\_91}: sup = 0.5, {gene\_59, gene\_6}: sup = 0.51

{gene\_59, gene\_72}: sup = 0.62, {gene\_59, gene\_87}: sup = 0.51

L3 - {gene\_5, gene\_1, gene\_3}: sup = 0.52

{gene\_59, gene\_1, gene\_72}: sup = 0.5

Length-3 Candidate Itemsets from Gene Dataset

C3 - [frozenset({'gene\_1', 'gene\_47', 'gene\_3'}), frozenset({'gene\_1', 'gene\_3', 'gene\_5'}), frozenset({'gene\_1', 'gene\_59', 'gene\_3'}), frozenset({'gene\_1', 'gene\_3', 'gene\_72'}), frozenset({'gene\_1', 'gene\_47', 'gene\_5'}), frozenset({'gene\_1', 'gene\_59', 'gene\_5'}), frozenset({'gene\_1', 'gene\_6', 'gene\_5'}), frozenset({'gene\_1', 'gene\_5', 'gene\_72'}), frozenset({'gene\_1', 'gene\_5', 'gene\_87'}), frozenset({'gene\_1', 'gene\_91', 'gene\_5'}), frozenset({'gene\_1', 'gene\_59', 'gene\_6'}), frozenset({'gene\_1', 'gene\_59', 'gene\_72'}), frozenset({'gene\_1', 'gene\_59', 'gene\_87'}), frozenset({'gene\_47', 'gene\_3', 'gene\_5'}), frozenset({'gene\_3', 'gene\_5', 'gene\_72'}), frozenset({'gene\_59', 'gene\_3', 'gene\_5'}), frozenset({'gene\_59', 'gene\_3', 'gene\_72'}), frozenset({'gene\_59', 'gene\_6', 'gene\_5'}), frozenset({'gene\_59', 'gene\_5', 'gene\_72'}), frozenset({'gene\_59', 'gene\_5', 'gene\_87'})]

Code for apriori\_gen function (printed C3 itemsets here to copy into this file since locating them was easier than from the print in main)

def apriori\_gen(freq\_sets, k):

    candidate\_list = []

    len\_freq\_sets = len(freq\_sets)

    for x in range(len\_freq\_sets):

        for y in range(x + 1, len\_freq\_sets):

            c\_items = list(freq\_sets[x])[:k-2]

            if c\_items == list(freq\_sets[y])[:k-2]:

                candidate\_itemset = freq\_sets[x] | freq\_sets[y]

                all\_subsets\_frequent = all(frozenset(subset) in freq\_sets for subset in itertools.combinations(candidate\_itemset, k-1))

                if all\_subsets\_frequent:

                    candidate\_list.append(candidate\_itemset)

    #print(candidate\_list)

    return candidate\_list

Code for get\_freq function

def get\_freq(dataset, candidates, min\_support, verbose=False):

    count = {}

    for transaction in dataset:

        for candidate in candidates:

            if candidate.issubset(transaction):

                count[candidate] = count.get(candidate, 0) + 1

    num\_transactions = float(len(dataset))

    freq\_list = []

    support\_data = {}

    for candidate in candidates:

        support = count[candidate] / num\_transactions

        if support >= min\_support:

            freq\_list.append(candidate)

        support\_data[candidate] = support

return freq\_list, support\_data