Title: What affects obesity

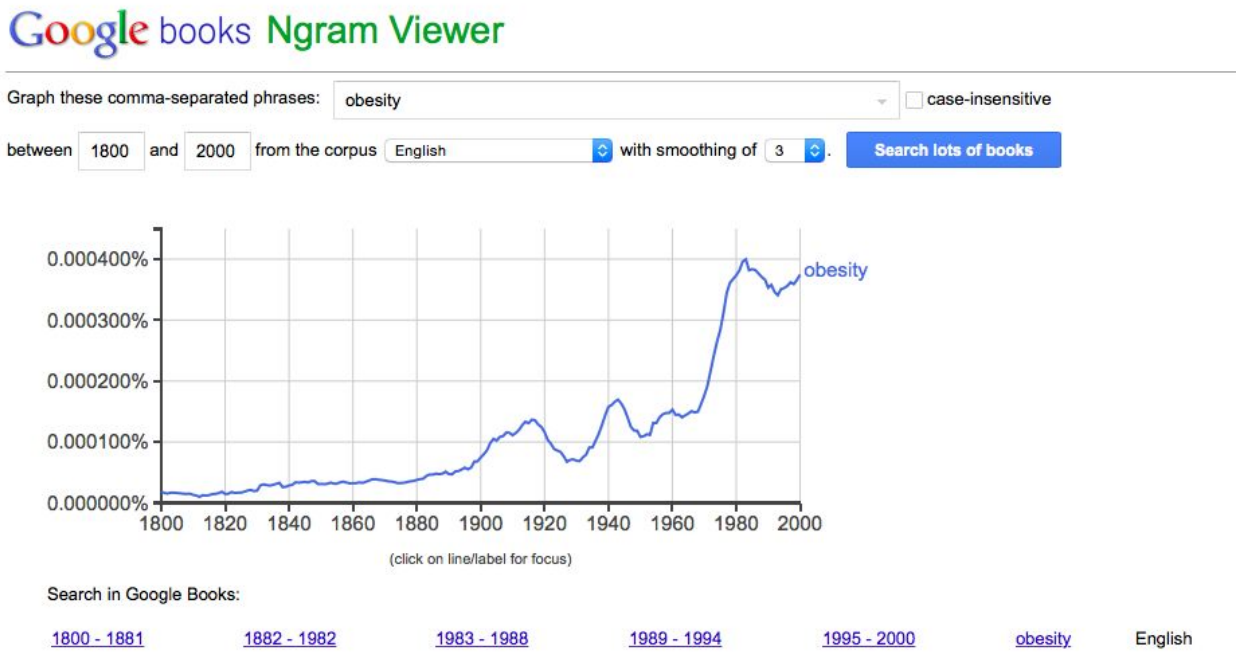
Team member: Jieying Chen
Yangyang Zhang
Yueyang Wu

Object and significance

Our goal for this project is to find the attributes affect the obesity in United States, and to find the weight of these attributes affect obesity problem. Trying to find a health lifestyle that reduces obesity people in United States. The reason why it important is the rate of obesity people in United State became higher and higher since 20th century no matter adult or child.

Obesity is a danger to health, which leads to a lot health issues, like diabetes, heart disease and so on. For adult, obesity is a enemy of a long and healthy life. The incidence of obesity complicated with cerebral embolism and heart failure 1 times higher than normal weight. Also, obesity people will increasing operation difficulty, and has higher risk of anesthesia, the wound after the operation is easy to crack. For children, The harm of childhood obesity is all aspects, to children's circulatory system Respiratory system, digestive system, endocrine system, immune system, such as multiple systems of the body cause serious damage, seriously affecting the healthy growth of children, children's intelligence development affects the development ability of mental health and sexual behavior. So observe and analyze attributes that impact obesity is an important and cannot be neglected task.

Background



From the database of all the books Google collected in English, it is obvious to see that obesity becomes a hot topic since 1980, and it is still a huge problem in the 20th century. It is also known that obesity can cause many other health problems even diseases, such as High Blood Pressure and cancer. That is the main reason that our group chose to finding out the connection between each factor and obesity. Another reason for us to choose this problem is that we think analysis the common factors for obesity will also be useful for finding out an effective way to prevent it. From the database of national obesity rate we found online, during 2011-2014, none of the fifty states of the United State has an obesity rate under 20%, which means that an extremely huge population is suffering from obesity, and many are also suffering the pain by diseases related to obesity.

We found some previous works online which discusses some factors of obesity, such as the "what causes overweight and obesity?" article we found from National Heart, Lung, and Blood Institute website. However, there's no article puts all the factors together and discuss the overall fact, which means to compare all the attributes we found and discover which one is the most likely one that will cause overweight.

Method

- Attributes:

The age group (adults, adolescents, children)

The affection of Environmental supports such as state's policy on P.E classes

The activities in different people do during leisure time

The food and nutrition that people eat that will greatly affect people's health.

Describe your data and how you obtained it.

we get our data from some government website and organization website which provide a lot data that useful and credible.

In the Center of Disease Control and Prevention website (CDC), the database it proved contain 52 states, and It sorting data by year from 2011 to present. Also, for each attribute, the data is concretely. We select National data to get the comprehensive view. Also, we choose the data from three age groups which are children, adolescents, and adults. When we analysis the Adults who consume fruit < 1 time daily nationally, we using a sample that contain 442,993 people in 2013. Similarly, for Adults who consume vegetables < 1 time daily, the sample size is 434,803 in 2013. For adolescents, we also analyzed these two factors, but sample size is smaller compare to adults. the sample sizes are 13,322 and 13,227. For Physical activity factors, we using a sample size which contain 450,093 people, and research adults who engage in no leisure time physical activity. For adolescents, we choose the drank soda daily and watching 3 or more hours of TV daily these two factors. The sample size for drank soda daily is 13,324, and the sample size for watching TV 3 or more hours is 13,245.

For Children, we using the data resources from organization website that for child and adolescent health. We choose the factor that the highest education of adult in household.

Why we select these attributes:

By seeing the results from the health status, we can easily see how the bad habits in life affect people's health.

Describe your methodology : Using flowcharts, diagram, or formula

We use the Apriori algorithm to find out the frequent subset of the collection of features which we considered might be the features that mostly affects obesity to occur.

Here's the pseudocode for the algorithm, where T is the database and ϵ is a support threshold.

$$
\begin{aligned}
&\text{Apriori}(T, \epsilon) \\
&\quad L_1 \leftarrow \{\text{large } 1-\text{itemsets}\} \\
&\quad k \leftarrow 2 \\
&\quad \textbf{while } L_{k-1} \neq \emptyset \\
&\quad\quad C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a\} - \{c \mid \{s \mid s \subseteq c \wedge |s| = k-1\} \nsubseteq L_{k-1}\} \\
&\quad\quad \textbf{for } \text{transactions } t \in T \\
&\quad\quad\quad C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\} \\
&\quad\quad\quad \textbf{for } \text{candidates } c \in C_t \\
&\quad\quad\quad\quad count[c] \leftarrow count[c] + 1 \\
&\quad\quad L_k \leftarrow \{c \mid c \in C_k \wedge \ count[c] \geq \epsilon\} \\
&\quad\quad k \leftarrow k + 1 \\
&\quad \textbf{return } \bigcup_k L_k
\end{aligned}
$$

In addition, here's the steps we use to implement the algorithm.

- First, we build a Candidate list of k-itemsets and then extract a frequent list of k-itemsets using the support count
- Then, we use the frequent list of k-itemsets in determining the candidate and frequent list of k+1 itemsets
- We use Pruning to do that
- We repeat until we have an empty candidate or frequent of k-itemsets. Then we return the list of k-1 itemsets
- Also, we calculated the support value for each subset. We set the minimum support to be 0.5 and only print out the subsets whose support is over the minimum support.

Describe evaluation strategy: how will you measure your performance of your algorithm.

We write the python code which can print out the frequent sets of features which may cause obesity of each age group, as well as the support of each subset. Then, we use matlab to draw the table for the output. And it is clearly to see that which combination of the features will have the largest support, which means it impacts most on causing obesity.

Questions we want to discover:

1. How does breastfeeding influencing obesity?
2. Among consuming less fruit and vegetable, having less leisure time, which feature affects adult obesity most?
3. Among consuming less fruit and vegetable, having less leisure time, watching too much TV, drinking too much soda, which feature affects adult obesity most?
4. How does parents' educational level affects child obesity?

Result: (2-5)

```
#-*- coding:utf-8 - *-

filename = "/Users/Derella/Desktop/Data/adults.txt"

def load_dataset():
    "Load the sample dataset."
    dataset = []
    with open(filename,'r') as target:
        for line in target:
            smalllist =[x.strip() for x in line.split(',')]
            dataset.append(smalllist)
    datalist = []
    list1 = [item[0] for item in dataset]
    list2 = [item[1] for item in dataset]
    list3 = [item[2] for item in dataset]

    average1 = sum(float(a) for a in list1)/len(list1)
    average2 = sum(float(a) for a in list2)/len(list2)
    average3 = sum(float(a) for a in list3)/len(list3)

    for b in dataset:
        smalllist = []
        if float(b[0]) > average1:
            smalllist.append(b[4])
        if float(b[1]) > average2:
            smalllist.append(b[5])
        if float(b[2]) > average3:
            smalllist.append(b[6])
        datalist.append(smalllist)
    datalist = filter(None,datalist)
    return datalist

def createC1(dataset):
    "Create a list of candidate item sets of size one."
    c1 = []
    for transaction in dataset:
        for item in transaction:
            if not [item] in c1:
                c1.append([item])
    c1.sort()
    #frozenset because it will be a ket of a dictionary.
    return map(frozenset, c1)


def scanD(dataset, candidates, min_support):
    "Returns all candidates that meets a minimum support level"
    sscnt = {}
    for tid in dataset:
        for can in candidates:
            if can.issubset(tid):
                sscnt.setdefault(can, 0)
                sscnt[can] += 1

    num_items = float(len(dataset))
    retlist = []
    support_data = {}
    for key in sscnt:
        support = sscnt[key] / num_items
        if support >= min_support:
            retlist.insert(0, key)
        support_data[key] = support
    return retlist, support_data
```

```
def aprioriGen(freq_sets, k):
    "Generate the joint transactions from candidate sets"
    retList = []
    lenLk = len(freq_sets)
    for i in range(lenLk):
        for j in range(i + 1, lenLk):
            L1 = list(freq_sets[i])[:k - 2]
            L2 = list(freq_sets[j])[:k - 2]
            L1.sort()
            L2.sort()
            if L1 == L2:
                retList.append(freq_sets[i] | freq_sets[j])
    return retList


def dumpclean(obj):
    if type(obj) == dict:
        for k, v in obj.items():
            if hasattr(v, '__iter__'):
                print k
                dumpclean(v)
            else:
                print '%s : %s' % (k, v)
    elif type(obj) == list:
        for v in obj:
            if hasattr(v, '__iter__'):
                dumpclean(v)
            else:
                print v
    else:
        print obj

def apriori(dataset, minsupport=0.5):
    "Generate a list of candidate item sets"
    C1 = createC1(dataset)
    D = map(set, dataset)
    L1, support_data = scanD(D, C1, minsupport)
    L = [L1]
    k = 2
    while (len(L[k - 2]) > 0):
        Ck = aprioriGen(L[k - 2], k)
        Lk, supK = scanD(D, Ck, minsupport)
        support_data.update(supK)
        L.append(Lk)
        k += 1
    print("subsets: ")
    print L
    print("supports: ")
    dumpclean(support_data)
    return
```

The code above is to implement Apriori algorithm to gather the subsets of the dataset we find. First we load the dataset file and modify the datalist we want to use. Then the createC1() function is be used to create all the combination of the sets based on the features we have. After that, we calculated the support for each candidate, which is the subset of the modified dataset. In the end, we apply the algorithm and print out the result we get. The function dumpclean() is only be used to make the printout results look nice.

T =

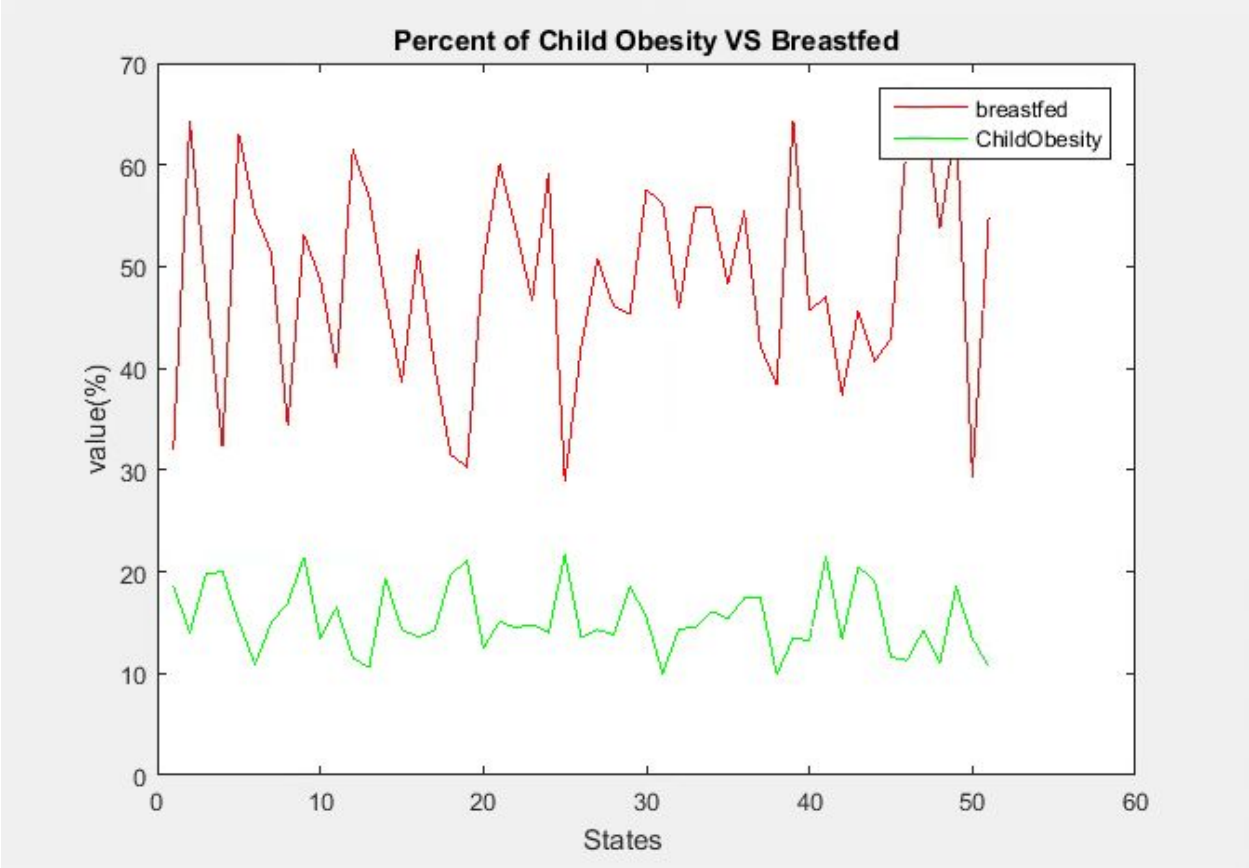
	adultsupprot
no leisure timeconsuming <1 vegetables dailyconsuming <1 fruits daily	0.44118
consuming <1 vegetables daily	0.73529
consuming <1 fruits daily	0.64706
no leisure timeconsuming <1 vegetables daily	0.5
consuming <1 vegetables dailyconsuming <1 fruits daily	0.5
no leisure time	0.70588
no leisure timeconsuming <1 fruits daily	0.52941

The table above is the result of the adult's obesity frequent sets and supports.

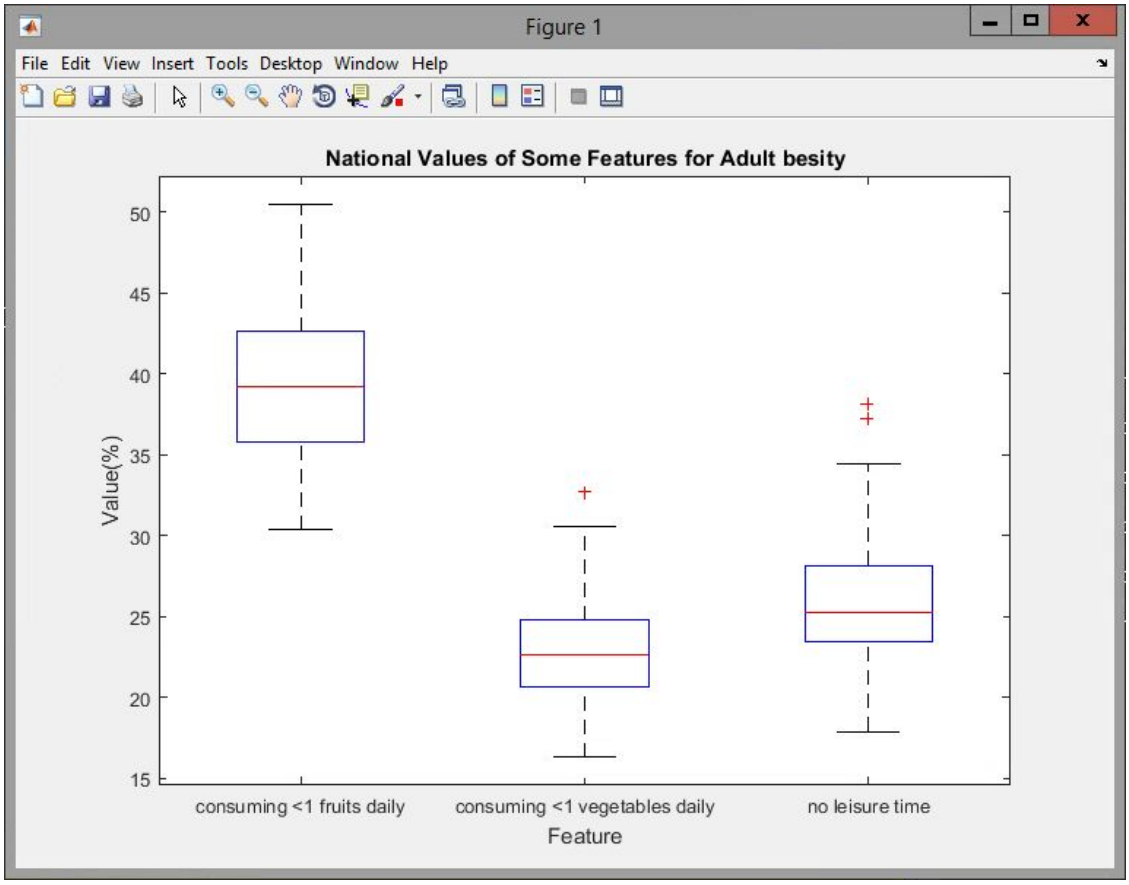
T2 =

	adolescent_support
consuming <1 fruits daily	0.35714
no leisure timeconsuming <1 vegetables dailywatch 3+ hrs TV daily	0.57143
consuming <1 vegetables dailywatch 3+ hrs TV daily	0.59524
consuming <1 vegetables daily	0.7619
drink soda at least 1 daily	0.38095
no leisure timeconsuming <1 vegetables daily	0.7381
watch 3+ hrs TV daily	0.7619
no leisure timewatch 3+ hrs TV daily	0.71429
no leisure time	0.95238

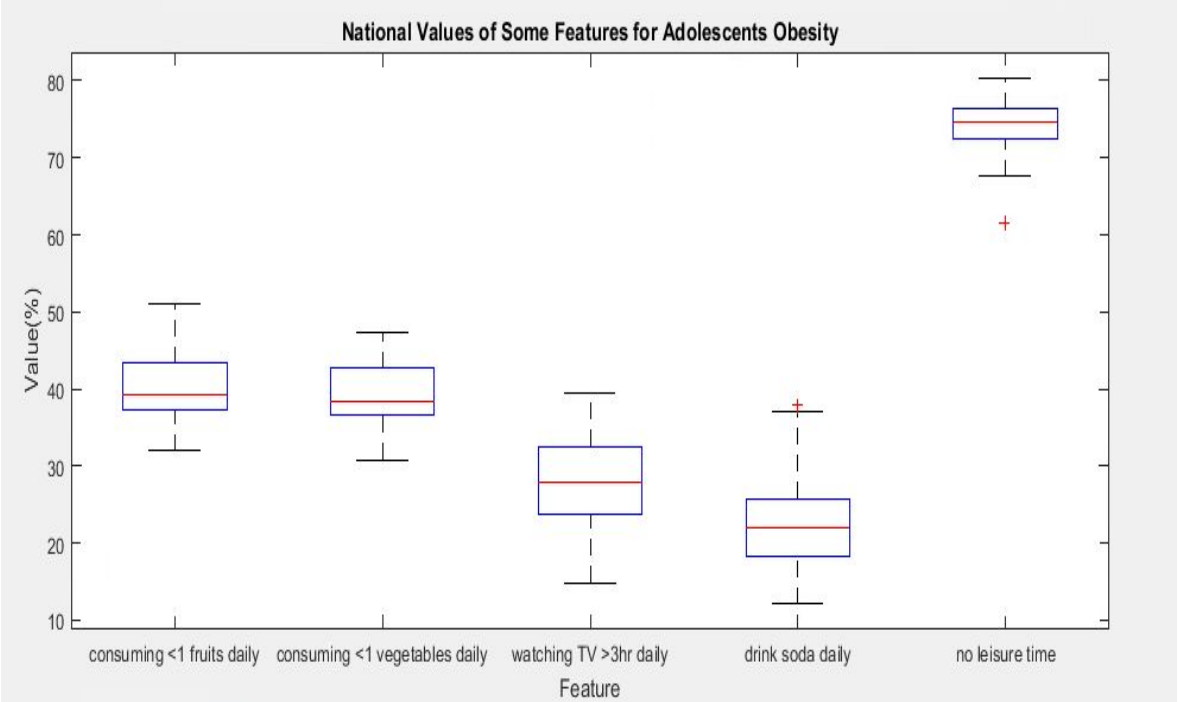
The table above is the result of the adolescent's obesity frequent sets and supports.



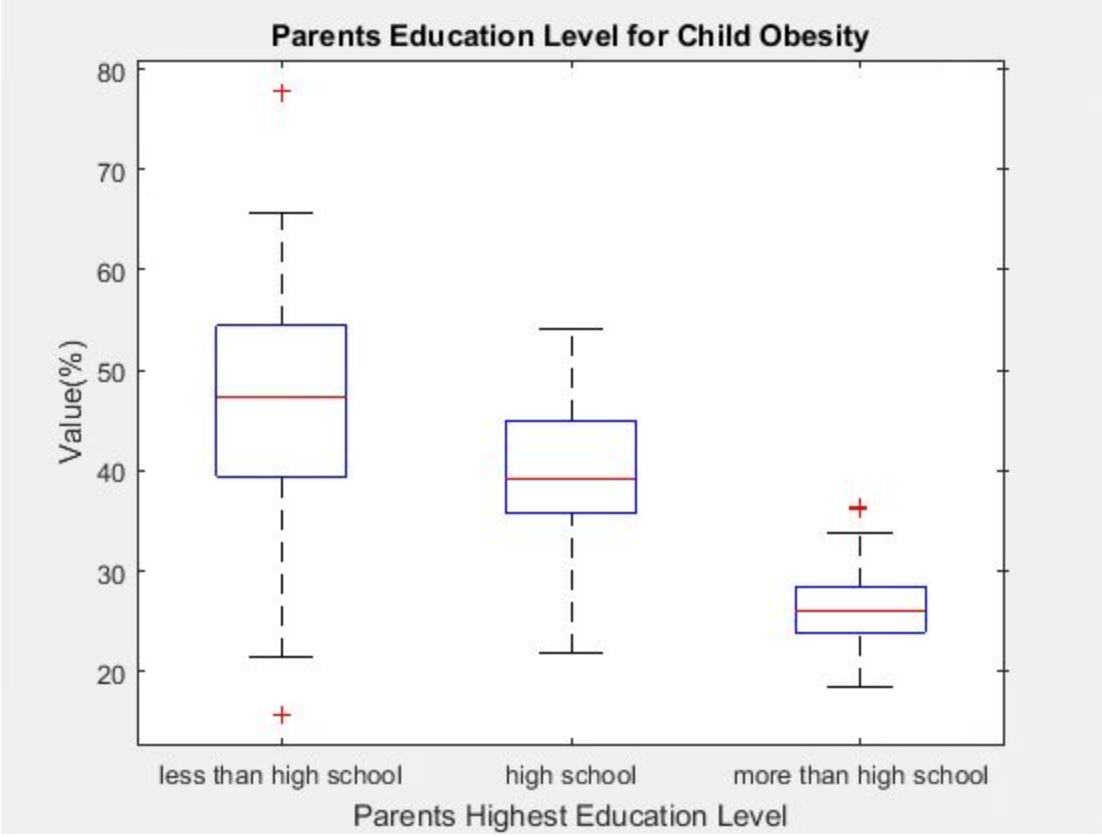
This table shows the obesity rate of children and the percent of breastfed child of data collected from year 2011. It is clear from the graph that the rate of breastfed child and the child obesity is inverse. And thus we can conclude that breastfeeding can have a positive effect on control the obesity rate of child.



This table is using Box plot to show some features affects Adult obesity. We have three features as horizontal axis which are people who consuming < 1 fruit daily, people who consuming < 1 vegetables daily and people who don't do physical activities because no leisure time, taking obesity rate as vertical axis. As the figure show, adult who don't consuming one fruit per day have higher obesity rate than people who don't do physical activities. Also, people who don't do physical activities has higher obesity rate than people who don't consuming one vegetable daily. But we also can see, among these three features, they all have really high obesity rate.



This table is using Box plot to show some features affects Adolescents obesity. We have five features as horizontal axis which are adolescents who consuming < 1 fruit daily, adolescents who consuming < 1 vegetables daily, adolescents who watching TV more than three hours daily, adolescents who drink soda more than one daily and adolescents who don't do physical activities because no leisure time, we taking obesity rate as vertical axis. As the figure show, adolescents who don't do physical activities during the leisure time have highest obesity rate, and it has really high rate which means this one feature has strong relationship with adolescents obesity. Also, consuming less than one fruit daily and consuming less than one vegetables daily, these two factors also causing really high obesity rate.



This table is using Box plot to explain how parents highest education level affects Child obesity. We have the different parents highest education as horizontal axis, and obesity rate as vertical axis. As the figure show, children whose parents' highest education level is less than high school have higher percent of obesity rate. and with parents education increase, the obesity rate of children will decrease.

Conclusion:(1page)

From the figure we get after we input the breastfed data and child obesity data in 2011. We found out that it is clear the when the breastfed rate went very low, the child obesity rate go very high. So, there is obviously a connection between child obesity and breastfeeding: when more people use breastfeeding, less children will be obesity.

We get the result of the frequent sets and supports over the minimum support which we set to be 0.5 of adult obesity and adolescents obesity. From the first table of adult obesity, we found out that among consuming less fruit and vegetable and having less leisure time, the feature that affects adult obesity most is consuming less than one vegetables daily. This is the result we expect to get before we do the experiment because not many families go to the grocery stores and markets very often. Then most of them will choose the frozen food instead of fresh vegetables because it is obvious that frozen food will last longer than fresh food.

For adolescent's obesity, we create a table based on consuming less fruit and vegetable, having less leisure time, watching too much TV and drinking too much soda. It would be surprising for many people to see that having no leisure time is the most effective factor. This may because that the original data we get is 'children and adolescents' rate on who are physically activated', then we subtracted that data using 100 to get the data of children and adolescents who are not physically activated.

When discovering the last question, we found out that children whose parents' highest education level is less than high school have higher percent of obesity rate. and with parents education increase, the obesity rate of children will decrease.

Individual Task:

Each team member of this team is important because everyone's idea is valuable and might be different on the same thing. It would be really useful to have several different opinions together, and each opinion will be a totally new way to approach the result we expected to.

Jieying Chen's main task is to analysis the tables for factors that causes adult's obesity, including environmental factors such as the state's policy about smoking. The databases we found also contains some information we don't need, so that every single member need to identify the part we need and store that information for further use. Yangyang Zhang's task is mostly focused on the adolescents. There are also several datasets that contains the factors that can cause adolescents' overweight. Yueyang Wu's task is to search for the databases that containing the children's and infants' obesity causes, such as the way of feeding and parents smoke or not.

For every single member, the shared tasks are to identify the issue, to work together as a team communicating well and analysis the final useful dataset that we selected from several databases together to get the result we expected.

Jieying - Code, adult

Yangyang - Adolescent, collect & describe data

Yueyang Wu - Infant, Find evaluation strategy

Together: Find result

Reference:

https://books.google.com/ngrams/graph?content=obesity&year_start=1800&year_end=2000&corpus=15&smoothing=3&share=&direct_url=t1%3B%2Cobesity%3B%2Cc0

What Causes Overweight and Obesity? (n.d.). Retrieved October 21, 2015, from http://www.nhlbi.nih.gov/health/health-topics/topics/obe/causes

National Location Summary. (n.d.). Retrieved October 21, 2015, from http://nccd.cdc.gov/NPAO_DTM/LocationSummary.aspx?statecode=94

http://www.cdc.gov/breastfeeding/data/reportcard.htm

http://www.childhealthdata.org/browse/allstates?q=2415&a=3879&g=470

http://www.ncsl.org/research/health/childhood-obesity-trends-state-rates.aspx#2003

Links might be useful for future exploration:

http://www.cdc.gov/physicalactivity/downloads/PA_State_Indicator_Report_2010.pdf

http://www.fitness.gov/resource-center/facts-and-statistics/

http://www.healthdata.gov/

http://www.fitness.gov/resource-center/facts-and-statistics/

http://subjectguides.library.american.edu/c.php?g=175023&p=1155321

http://archive.ics.uci.edu/ml/

https://catalog.data.gov/dataset?_organization_limit=0&organization=hhs-gov#topic=health_navigation

http://nccd.cdc.gov/NPAO_DTM/LocationSummary.aspx?state=Indiana

http://catalog.data.gov/dataset/community-health-status-indicators-chsi-to-combat-obesity-heart-disease-and-cancer

http://jama.jamanetwork.com/article.aspx?articleid=1832542