In [1]:
```
%load_ext sql
%sql mysql+pymysql://root:@fe512_mysql/fe512db
%sql USE hint
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
0 rows affected.
```

Out[1]: []

FE512 Final Project

MySQL Database Design for Survey Data on Cancer

& Cancer Risk Factors Analysis

Group: 14
Authors: Tianrui Wang & Xiaojun Zhu
Instructor: Prof. Olorundamilola Kazeem
TA: Tingyi Lu
Date: May 20, 2019

INTRODUCTION

Cancer has a huge impact on society across the world. Cancer statistics help us understand what happens in large groups of people and provide a big picture of the burden of cancer over all times (National Cancer Institution, 2018). This project makes use of the Health Information National Trends Survey (HINTS) to derive information such as how many people are diagnosed with cancer in a given year, the average age at diagnosis, and differences among groups defined by interested categories.

The HINTS data collection program was created by the National Cancer Institute (NCI) to monitor changes in the rapidly evolving field of health communication. It collects nationally representative data routinely about the American public's use of cancer-related information.

This project aims to build an easy-to-use SQL database to store and use the data collected by the Health Information National Trends Survey (HINTS). In the process, we learned the structure of the survey data, modified and broke down the original dataset with over 300 variables, and developed a strategy for MySQL database design. In the end, the project also investigated several leading risk factors through MySQL queries for all types of cancers and especially for prostate cancer.

Hopefully, the project could help the audience understand the myths and true risk factors, and provide information to non-professionals in cancer care management.

DATA SOURCE

The project uses the 2017 Cycle 1 data from HINTS by National Cancer Institution (https://hints.cancer.gov/data/download-data.aspx (https://hints.cancer.gov/data/download-data.aspx)). The following explanation is qouted from National Cancer Institution for readers to understand the data content:

```
"HINTS collects data about the use of cancer-related information by the Amer
ican public. These data provide opportunities to understand and improve heal
th communication.

* Provides updates on changing patterns, needs, and information opportunitie
s in health
* Identifies changing communications trends and practices
* Assesses cancer information access and usage
* Provides information about how cancer risks are perceived
* Offers a testbed to researchers to test new theories in health communicati
on"
```

DATA MODEL

1. Data Structure of the Original Data

The HINT Survey is a set of survey questions broken down by sections with different focuses, for example, "Looking For Health Information", "Using the Internet to Find Information", "Your Health Care". There are total 15 sections coded from A to O.

Each sections contain a few "True/False" questions, multiple choices questions, or multiple choices questions with a blank field to specify unmentioned choices. The answer to each questions are coded as differente categorical variables with number values representing different categories, e.g., 1 for Yes, 2 for No, -9 for missing data, etc.

The dataset provided by NCI is a set with over 3000 rows and more than 300 variables.

Figure 1: A True/False Question

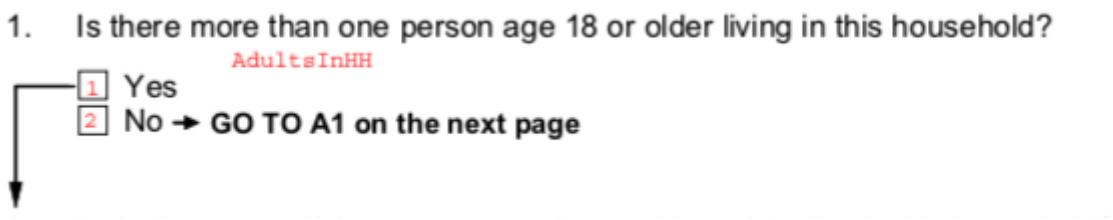


Figure 2: A Multiple choices Question

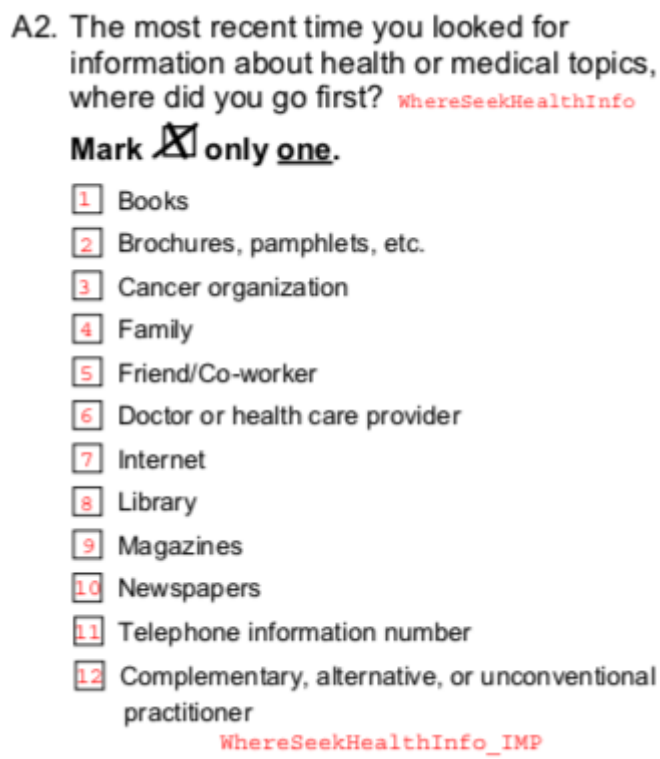


Figure 3: A Multiple choices Question with a blank field to specify unmentioned choices

A7. Imagine that you had a strong need to get information about health or medical topics. Where would you go first?

StrongNeedHealthInfo

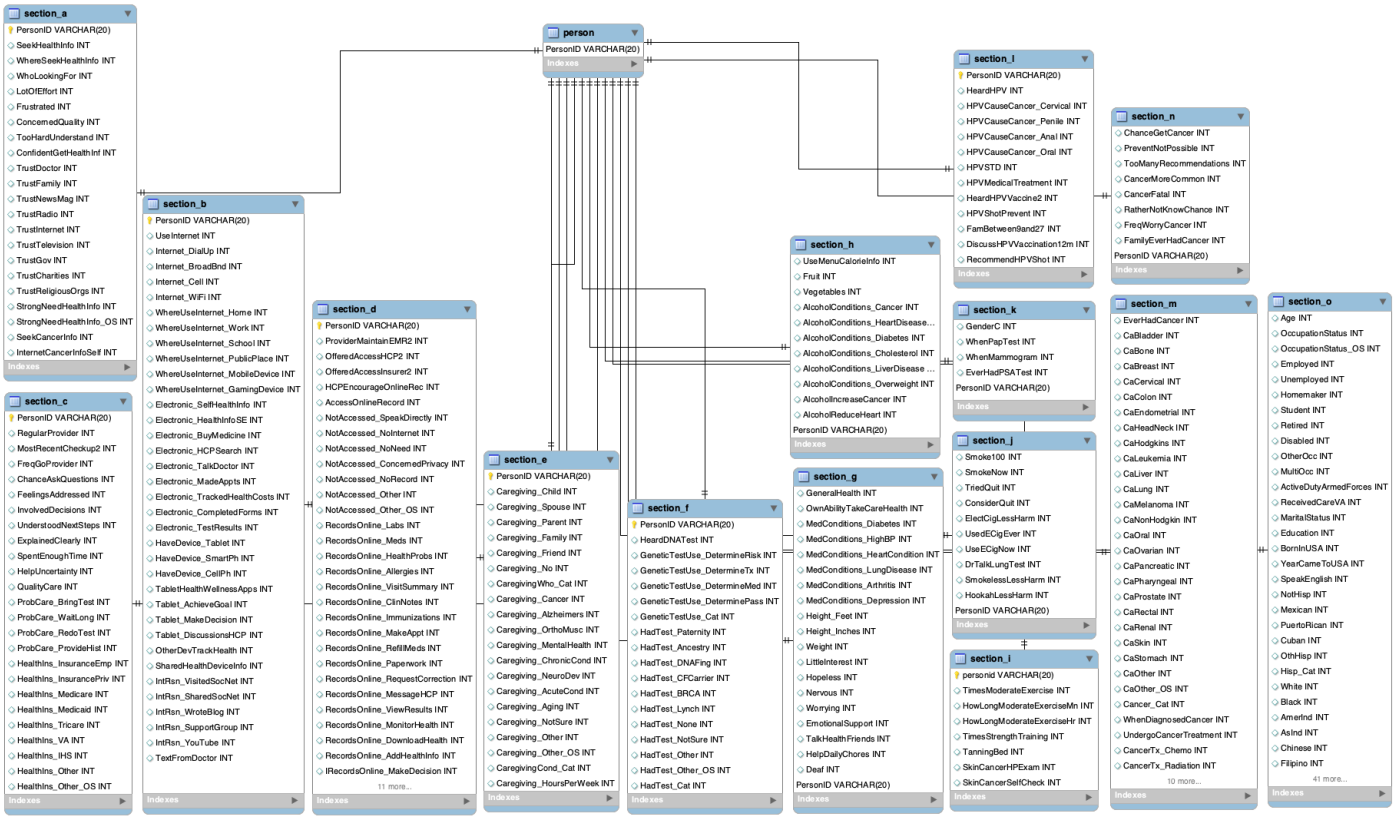
Mark ☒ only one.

- [1] Books
- [2] Brochures, pamphlets, etc.
- [3] Cancer organization
- [4] Family
- [5] Friend/Co-worker
- [6] Doctor or health care provider
- [7] Internet
- [8] Library
- [9] Magazines
- [10] Newspapers
- [11] Telephone information number
- [12] Complementary, alternative, or unconventional practitioner
- [91] Other-Specify → StrongNeedHealthInfo_OS

2. ER Diagram of the Basic Database

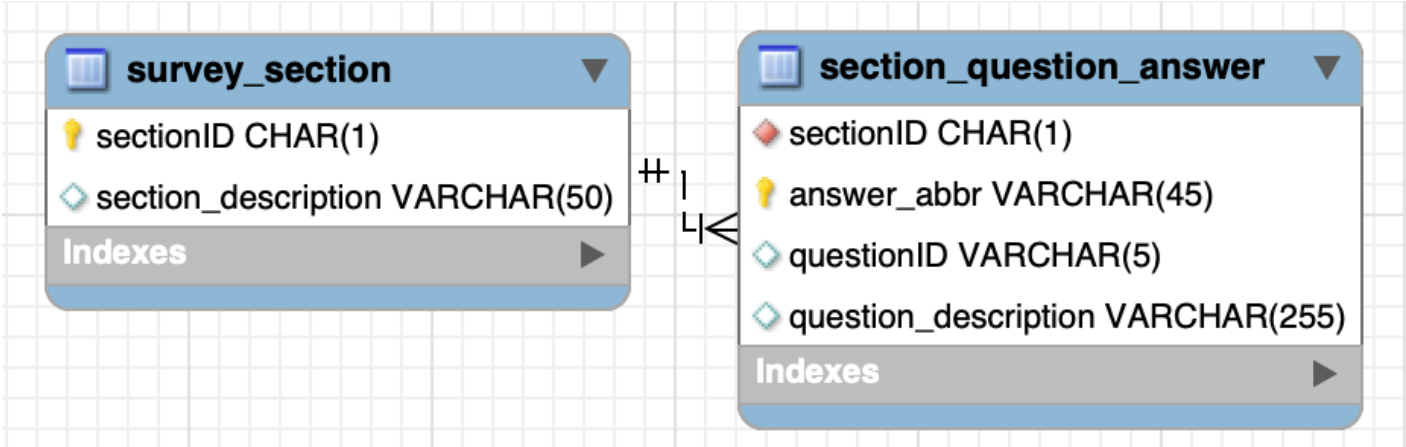
For easier handling of the dataset, we decomposed the original data into sections a to o horizontally, and created 15 tables accordingly. After decomposition, each table contains less than 72 variables. In each section table, 'PersonID' works as primary keys. Another table, 'person', helds all the unique IDs for each respondents and connects all the section tables.

Figure 4: ER Diagram Part I



We also constructed two tables that held the meta data, to help the database users understand and use the data. Inside the 'survey_section' table, two columns lists the sectionIDs and the descriptions for each section. The 'section_question_answer' table lists the variable names inside main tables as 'answer_abbr', the related question IDs and section IDs, as well as descriptions for each variable. These two tables are connected through the section IDs.

Figure 5: ER Diagram Part II



3. ER Diagram of Tables for Analysis

- Create tables and import data for each section from section_a to section_o
- Select desired variables and create new tables for analysis:
  - TABLE cancer_info - General Analysis
  - TABLE prostate_analysis - Prostate Cancer Analysis

As our analysis may require variables from different sections, we established two seperate tables for general and prostate cancer analysis.

3.1 General Cancer Analysis

Cancer is related to many factors, people habitat, people disease record, people attitude towards healthy, and basic informations. The factors we are interested in are listed in Figure 6:

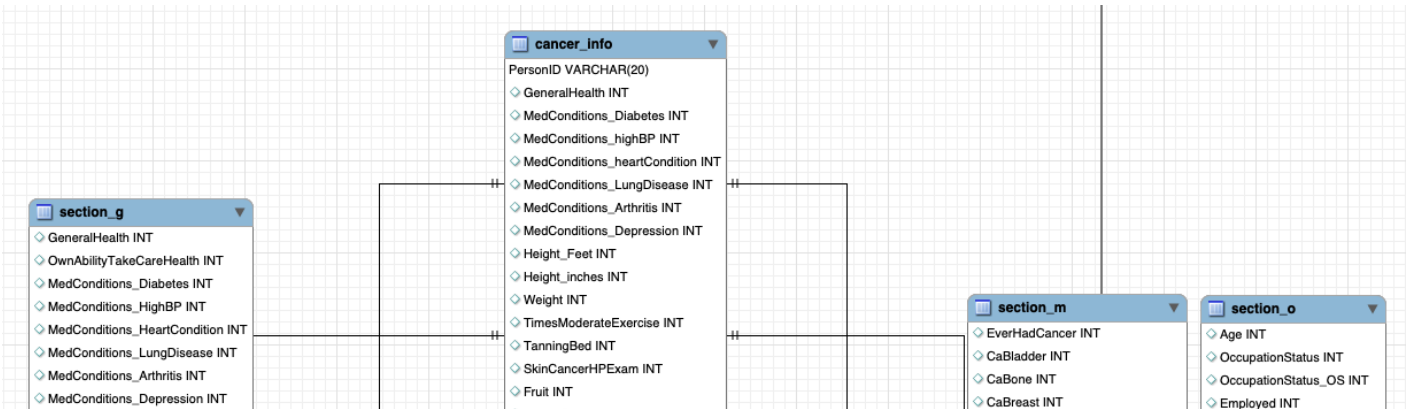
Figure 6: Selected factors for General Cancer Analysis

Table 1 listed the details of the selected categorical variables, their values, and explanation for their values. All of the information is obtained through the codebook by HINTS(NCI, 2017).

Column Name	Description	Value	Table Name
MedConditions_Diabetes	healthy history_Diabetes or high blood sugar?	1-2	Section_G
MedConditions_HighBP	healthy history_High blood pressure or hypertension?	1-2	Section_G
MedConditions_HeartCondition	healthy history_A heart condition such as heart attack, angina, or congestive heart failure?	1-2	Section_G
MedConditions_LungDisease	healthy history_Chronic lung disease, asthma, emphysema, or chronic bronchitis?	1	Section_G
MedConditions_Arthritis	healthy history_Arthritis or rheumatism?	1-2	Section_G
MedConditions_Depression	healthy history_Depression or anxiety disorder?	1-2	Section_G
Height_Feet	height in feet	number	Section_G
Height_Inches	height in inches	number	Section_G
Weight	weight in pounds	number	Section_G
Fruit	About how many cups of fruit (including 100% pure fruit juice) do you eat or drink each day?	1-6	Section_H
Vegetables	About how many cups of vegetables (including 100% pure vegetable juice) do you eat or drink each day?	1-7	Section_H
TimesModerateExercise	In a typical week, how many days do you do any physical activity or exercise of at least moderate intensity, such as brisk walking, bicycling at a regular pace, and swimming at a regular pace?	0-7	Section_I

Column Name	Description	Value	Table Name
TanningBed	How many times in the past 12 months have you used a tanning bed or booth?	0-4	Section_I
SkinCancerHPExam	Do you ever have your skin examined by a health professional for signs of skin cancer?	1-4	Section_I
Smoke100	Have you smoked at least 100 cigarettes in your entire life?	1-2	Section_J
Gender	Are you male or female?	1-2	Section_K
EverHadCancer	Have you ever been diagnosed as having cancer?	1-2	Section_M
FamilyEverHadCancer	Have any of your family members ever had cancer?	1-2-4	Section_N
Age	What is your age?	number	Section_O
BornInUSA	Were you born in the United States?	1-2	Section_O
IncomeRanges	what is your combined annual income,	1-9	Section_O
GeneralHealth	In general, would you say your health is...	1-5	Section_G

Figure 7: ER Diagram Part III

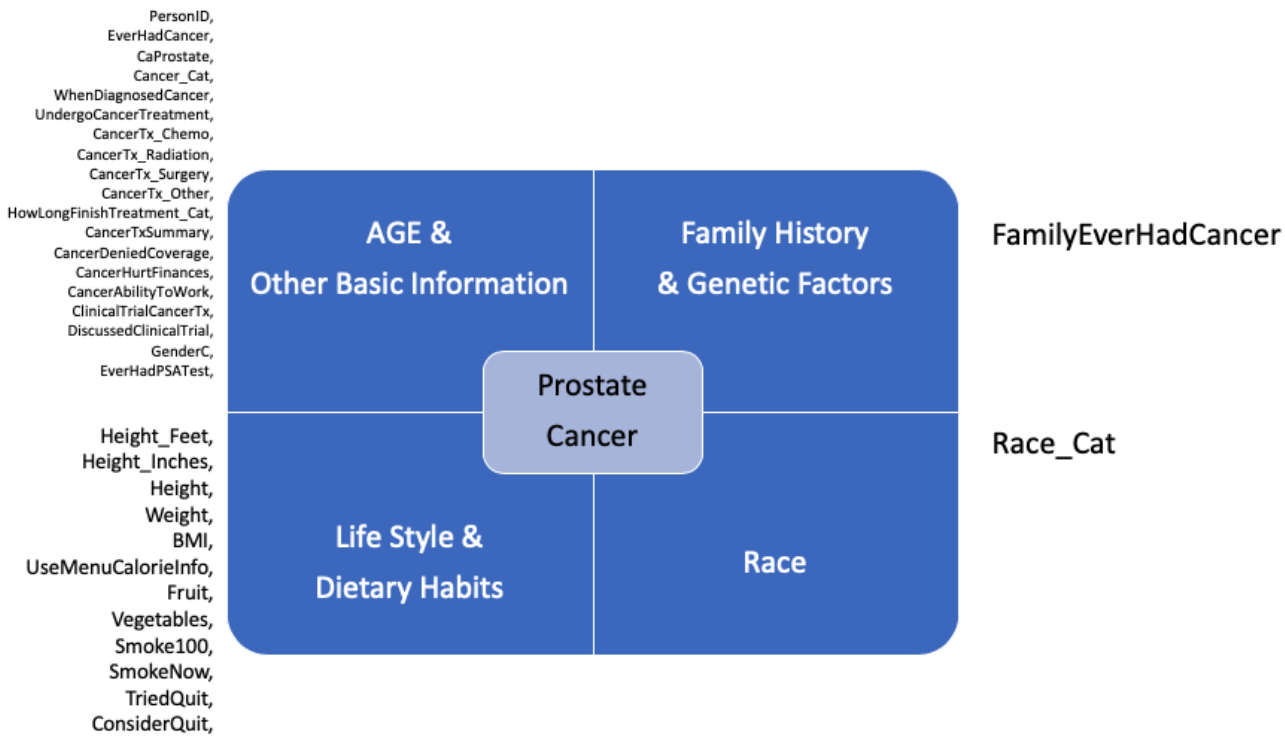

cancer_info
PersonID VARCHAR(20)
GeneralHealth INT
MedConditions_Diabetes INT
MedConditions_highBP INT
MedConditions_heartCondition INT
MedConditions_LungDisease INT
MedConditions_Arthritis INT
MedConditions_Depression INT
Height_Feet INT
Height_inches INT
Weight INT
TimesModerateExercise INT
TanningBed INT
SkinCancerHPExam INT
Fruit INT
section_g
GeneralHealth INT
OwnAbilityTakeCareHealth INT
MedConditions_Diabetes INT
MedConditions_HighBP INT
MedConditions_HeartCondition INT
MedConditions_LungDisease INT
MedConditions_Arthritis INT
MedConditions_Depression INT
section_m
EverHadCancer INT
CaBladder INT
CaBone INT
CaBreast INT
section_o
Age INT
OccupationStatus INT
OccupationStatus_OS INT
Employed INT

3.2 Prostate Cancer Analysis

- Age/ Diagnosed Age
- Family history / Genetic factors
- Race
- Lifestyle / Dietary habits - Weight & Smoking Factors

Prostate cancer is also related to many factors. Among all, age is the biggest contribution to the cancer. People with a family history of prostate cancer are also generically more likely to get cancer than those without. Apart from these two, it is believed that men of Africa has a higher chance of getting prostate cancer than people of other races. Apart from this, we also wanted to exam the effect of weight and smoking contributing to prostate cancer. Figure 8 shows all the variables we selected for the analysis.

Figure 8: Selected Factors and Variables for Prostate Cancer Analysis



When constructing the ER Diagram, we noticed that there are two variables 'Cancer_Cat', and 'Race_Cat2', are categorical variables with more than 3 categories. We want to assign literal names to each numerical representation. Therefore, we inner-joined the main table with Table 2 and Table 3 to create two categorical variables with verbal names. The completed ER diagram is shown in Figure 9.

Cancer_Cat: a derived variable to categorize responses given to question M2, its value labels are given as follows:

Cancer_Cat	Value_Label
-9	Missing data (Not Ascertained)
-6	Missing data (Filter Missing)
-2	Question answered in error (Commission Error)

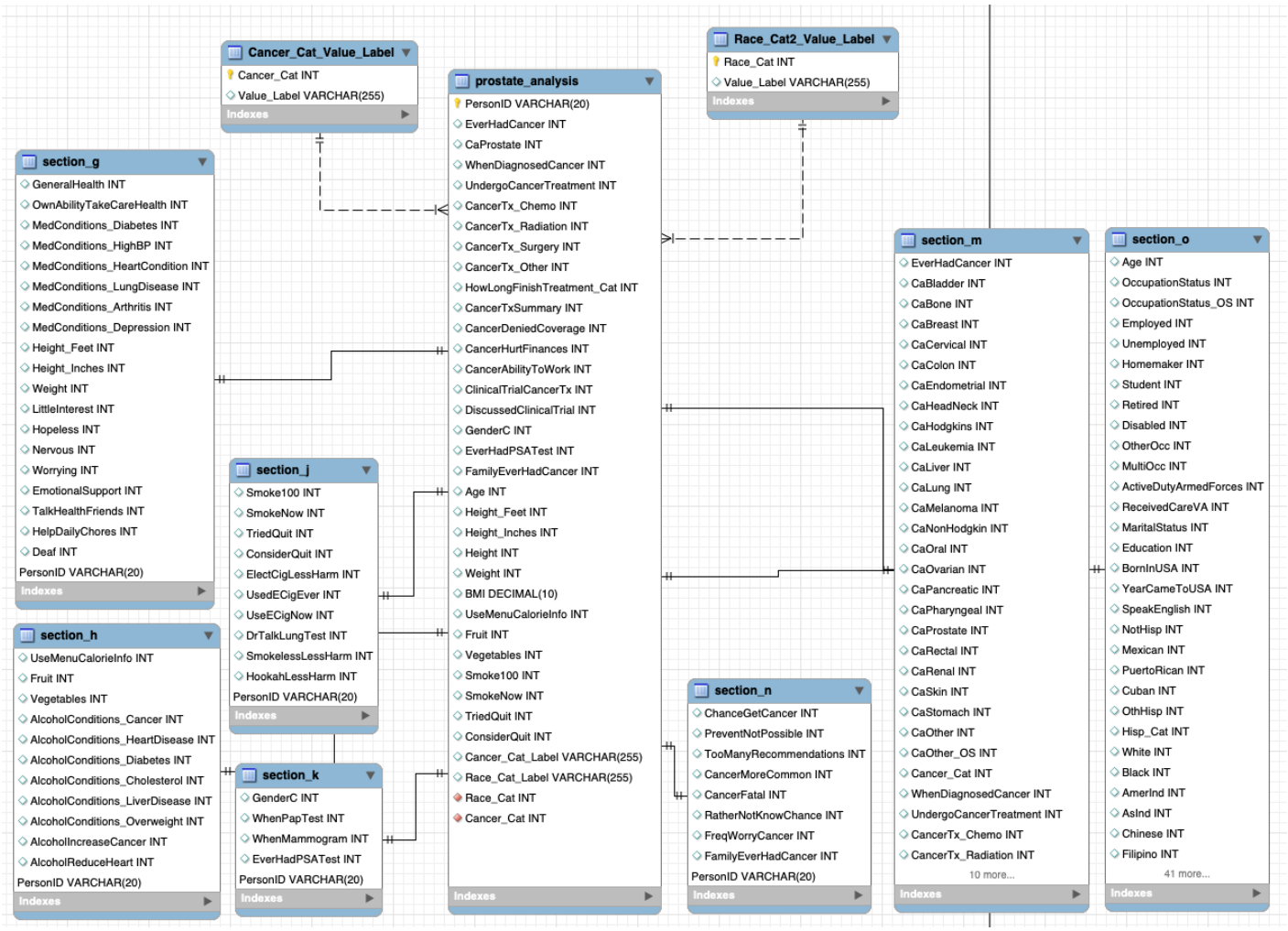
Cancer_Cat	Value_Label
-1	Inapplicable, coded 2 in EverHadCancer
1	Bladder cancer only
2	Bone cancer only
3	Breast cancer only
4	Cervical cancer only
5	Colon cancer only
6	Endometrial cancer only
7	Head/Neck cancer only
8	Hodgkins
9	Renal cancer only
10	Leukemia
11	Liver cancer only
12	Lung cancer only
13	Melanoma
14	Non-Hodgkin
16	Ovarian cancer only
17	Pancreatic cancer only
19	Prostate cancer only
20	Rectal cancer only
22	Skin cancer only
23	Stomach cancer only
25	More than one cancer checked
91	Other cancer only

Race_Cat2: a derived variable to categorize responses given in O11 (Race). The labels are given as follows:

Race_Cat2	Value Label
-9	Missing data (Not Ascertained)
11	White
12	Black
14	American Indian or Alaska Native
16	Multiple races selected
31	Asian Indian

Race_Cat2	Value Label
32	Chinese
33	Filipino
34	Japanese
35	Korean
36	Vietnamese
37	Other Asian
52	Guamanian or Chamorro
54	Other Pacific Islander

Figure 9: ER Diagram Part IV



4. Data Import & MySQL codes

We created and imported all the tables from Section A to Section O, the 'survey_section' table, and the 'section_question_survey' table. Appendix 1 & 2 are the jupyter notebooks that recorded the SQL code for the importing process.

Below are the codes for creating our tables for analysis.

4.1 Create Table 'Cancer_info'

In [2]:
```
%%sql
# According to above data strcture, we use foreign key"PersonaID" to joi
n columns from seperate tables into one table, we called "cancer_info"
# Set PersonID as primary key in table cancer_info
# Descripe table cancer_info
DROP TABLE IF EXISTS cancer_info;
CREATE TABLE cancer_info
    SELECT
        section_g.PersonID,
        section_g.GeneralHealth,
        section_g.MedConditions_Diabetes,
        section_g.MedConditions_highBP,
        section_g.MedConditions_heartCondition,
        section_g.MedConditions_LungDisease,
        section_g.MedConditions_Arthritis,
        section_g.MedConditions_Depression,
        section_g.Height_Feet,
        section_g.Height_inches,
        section_g.Weight,

        section_i.TimesModerateExercise,
        section_i.TanningBed,
        section_i.SkinCancerHPExam,

        section_h.Fruit,
        section_h.Vegetables,

        section_j.Smoke100,
        section_j.SmokeNow,

        section_k.GenderC,

        section_m.EverHadCancer,

        section_n.FamilyEverHadCancer,

        section_o.Age,
        section_o.BornInUSA,
        section_o.IncomeRanges
    FROM
        section_g,
        section_i,
        section_h,
        section_j,
        section_k,
        section_m,
        section_n,
        section_o
    WHERE(
        section_g.PersonID = section_i.PersonID  AND
        section_g.PersonID = section_h.PersonID  AND
        section_g.PersonID = section_j.PersonID  AND
        section_g.PersonID = section_k.PersonID  AND
        section_g.PersonID = section_m.PersonID  AND
        section_g.PersonID = section_n.PersonID  AND
        section_g.PersonID = section_o.PersonID
```

```
      )
;
ALTER TABLE cancer_info
  ADD PRIMARY KEY (PersonID);
DESCRIBE cancer_info;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
0 rows affected.
3285 rows affected.
0 rows affected.
24 rows affected.
```

Out[2]:

Field	Type	Null	Key	Default	Extra
PersonID	varchar(20)	NO	PRI	None	
GeneralHealth	int(11)	YES		None	
MedConditions_Diabetes	int(11)	YES		None	
MedConditions_highBP	int(11)	YES		None	
MedConditions_heartCondition	int(11)	YES		None	
MedConditions_LungDisease	int(11)	YES		None	
MedConditions_Arthritis	int(11)	YES		None	
MedConditions_Depression	int(11)	YES		None	
Height_Feet	int(11)	YES		None	
Height_inches	int(11)	YES		None	
Weight	int(11)	YES		None	
TimesModerateExercise	int(11)	YES		None	
TanningBed	int(11)	YES		None	
SkinCancerHPExam	int(11)	YES		None	
Fruit	int(11)	YES		None	
Vegetables	int(11)	YES		None	
Smoke100	int(11)	YES		None	
SmokeNow	int(11)	YES		None	
GenderC	int(11)	YES		None	
EverHadCancer	int(11)	YES		None	
FamilyEverHadCancer	int(11)	YES		None	
Age	int(11)	YES		None	
BornInUSA	int(11)	YES		None	
IncomeRanges	int(11)	YES		None	

4.2 Create Table 'prostate_analysis'

In [3]:
```
%%sql
DROP TABLE IF EXISTS prostate_analysis;
CREATE TABLE prostate_analysis
    SELECT
        section_m.PersonID,
        section_m.EverHadCancer,
        section_m.CaProstate,
        section_m.Cancer_Cat,
        #Cancer_Cat_Value_Label.Value_Lable AS Cancer_Cat_label,

        section_m.WhenDiagnosedCancer,
        section_m.UndergoCancerTreatment,
        section_m.CancerTx_Chemo,
        section_m.CancerTx_Radiation,
        section_m.CancerTx_Surgery,
        section_m.CancerTx_Other,  # cancer treatment
        section_m.HowLongFinishTreatment_Cat,
        section_m.CancerTxSummary,
        section_m.CancerDeniedCoverage,
        section_m.CancerHurtFinances,
        section_m.CancerAbilityToWork,
        section_m.ClinicalTrialCancerTx,
        section_m.DiscussedClinicalTrial,

        section_k.GenderC,
        section_k.EverHadPSATest,

        section_n.FamilyEverHadCancer, # family history

        section_o.Age, # age
        section_o.Race_Cat2, # race

        section_g.Height_Feet,
        section_g.Height_Inches,
        (section_g.Height_Feet * 12 + section_g.Height_Inches) AS Heigh
t, ##BMI
        section_g.Weight, ##BMI
        (section_g.Weight*703)/ ( POWER((section_g.Height_Feet * 12 + se
ction_g.Height_Inches) , 2) ) AS BMI,

        section_h.UseMenuCalorieInfo,
        section_h.Fruit,
        section_h.Vegetables,

        section_j.Smoke100,
        section_j.SmokeNow,
        section_j.TriedQuit,
        section_j.ConsiderQuit


    FROM
        section_m,
        section_k,
        section_n,
        section_o,
        section_g,
```

```
        section_h,
        section_j
#         section_m INNER JOIN Cancer_Cat_Value_Label
#         ON section_m.Cancer_Cat = Cancer_Cat_Value_Label.Cancer_Cat
    WHERE (
        # section_m.CaProstate = 1 AND
        section_m.PersonID = section_k.PersonID AND
        section_m.PersonID = section_n.PersonID AND
        section_m.PersonID = section_o.PersonID AND
        section_m.PersonID = section_g.PersonID AND
        section_m.PersonID = section_h.PersonID AND
        section_m.PersonID = section_j.PersonID #AND
        #section_m.Cancer_Cat = Cancer_Cat_Value_Label.Cancer_Cat
    )
;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
0 rows affected.
3285 rows affected.
```

Out[3]: []

In [4]:
```
%sql SELECT * FROM prostate_analysis LIMIT 1;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
1 rows affected.
```

Out[4]:

PersonID	EverHadCancer	CaProstate	Cancer_Cat	WhenDiagnosedCancer	UndergoCa
60000001-02	2	-1	-1	-1	-1

In [5]:
```
%%sql

# 1. Create & Import the Cancer_Cat_Value_Label.csv
# 2. Update the table prostate_analysis with a new column cotaining the
 Cancer_Cat_Label from a inner join of the two tables.
DROP TABLE IF EXISTS Cancer_Cat_Value_Label;
CREATE TABLE Cancer_Cat_Value_Label (
    Cancer_Cat INT,
    Value_Label VARCHAR(255)
);
LOAD DATA
    INFILE '/home/data/Cancer_Cat_Value_Label.csv'
    INTO TABLE Cancer_Cat_Value_Label
    FIELDS
        TERMINATED BY ','
        OPTIONALLY ENCLOSED BY '"'
        ESCAPED BY '\\'
    LINES
        TERMINATED BY '\r\n'
        STARTING BY ''
    IGNORE 1 LINES;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
0 rows affected.
0 rows affected.
26 rows affected.
```

Out[5]: []

In [6]:
```
%sql SELECT * FROM Cancer_Cat_Value_Label LIMIT 5;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
5 rows affected.
```

Out[6]:

Cancer_Cat	Value_Label
-9	Missing data (Not Ascertained)
-6	Missing data (Filter Missing)
-2	Question answered in error (Commission Error)
-1	Inapplicable, coded 2 in EverHadCancer
1	Bladder cancer only

In [7]:
```
%%sql
ALTER TABLE prostate_analysis
ADD Cancer_Cat_Label VARCHAR(255);
UPDATE prostate_analysis INNER JOIN Cancer_Cat_Value_Label
    ON prostate_analysis.Cancer_Cat = Cancer_Cat_Value_Label.Cancer_Cat
SET prostate_analysis.Cancer_Cat_Label = Cancer_Cat_Value_Label.Value_La
bel;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
0 rows affected.
3285 rows affected.
```

Out[7]: []

In [8]:
```
%sql SELECT * FROM prostate_analysis LIMIT 1;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
1 rows affected.
```

Out[8]:

PersonID	EverHadCancer	CaProstate	Cancer_Cat	WhenDiagnosedCancer	UndergoCa
60000001-02	2	-1	-1	-1	-1

In [9]:
```
%%sql
DROP TABLE IF EXISTS Race_Cat2_Value_Label;
CREATE TABLE Race_Cat2_Value_Label (
    Race_Cat INT,
    Value_Label VARCHAR(255)
);
LOAD DATA
    INFILE '/home/data/Race_Cat2_Value_Label.csv'
    INTO TABLE Race_Cat2_Value_Label
    FIELDS
        TERMINATED BY ','
        OPTIONALLY ENCLOSED BY '"'
        ESCAPED BY '\\'
    LINES
        TERMINATED BY '\r\n'
        STARTING BY ''
    IGNORE 1 LINES;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
0 rows affected.
0 rows affected.
14 rows affected.
```

Out[9]: []

In [10]:
```
%%sql
ALTER TABLE prostate_analysis
ADD Race_Cat_Label VARCHAR(255);
UPDATE prostate_analysis INNER JOIN Race_Cat2_Value_Label
    ON prostate_analysis.Race_Cat2 = Race_Cat2_Value_Label.Race_Cat
SET prostate_analysis.Race_Cat_Label = Race_Cat2_Value_Label.Value_Labe
l;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
0 rows affected.
3285 rows affected.
```

Out[10]: []

In [11]:
```
%sql SELECT * FROM prostate_analysis LIMIT 10;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
10 rows affected.
```

Out[11]:

PersonID	EverHadCancer	CaProstate	Cancer_Cat	WhenDiagnosedCancer	UndergoCa
60000001-02	2	-1	-1	-1	-1
60000006-02	2	-1	-1	-1	-1
60000011-01	2	-1	-1	-1	-1
60000014-01	2	-1	-1	-1	-1
60000017-01	2	-1	-1	-1	-1
60000019-01	2	-1	-1	-1	-1
60000020-01	2	-1	-1	-1	-1
60000022-01	2	-1	-1	-1	-1
60000025-02	1	2	22	60	1
60000026-01	2	-1	-1	-1	-1

DATA SUMMARY

Our data contains 3285 responses collected by HINTS in 2017. Since our data is stored in 15 tables by section, it is more valuable to look at a sample from the analysis tables 'cancer_info' & 'prostate_analysis'. Below are the queries for a sample of data.

In [12]:
```
%sql SELECT * FROM cancer_info LIMIT 10;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
10 rows affected.
```

Out[12]:

PersonID	GeneralHealth	MedConditions_Diabetes	MedConditions_highBP	MedConditi
60000001-02	2	2	1	2
60000006-02	4	2	1	2
60000011-01	3	2	2	2
60000014-01	4	1	1	1
60000017-01	2	2	2	2
60000019-01	1	2	2	2
60000020-01	3	2	1	2
60000022-01	2	2	2	2
60000025-02	3	2	1	1
60000026-01	2	2	2	2

In [13]:
```
%sql SELECT * FROM prostate_analysis LIMIT 10;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
10 rows affected.
```

Out[13]:

PersonID	EverHadCancer	CaProstate	Cancer_Cat	WhenDiagnosedCancer	UndergoCa
60000001-02	2	-1	-1	-1	-1
60000006-02	2	-1	-1	-1	-1
60000011-01	2	-1	-1	-1	-1
60000014-01	2	-1	-1	-1	-1
60000017-01	2	-1	-1	-1	-1
60000019-01	2	-1	-1	-1	-1
60000020-01	2	-1	-1	-1	-1
60000022-01	2	-1	-1	-1	-1
60000025-02	1	2	22	60	1
60000026-01	2	-1	-1	-1	-1

General Cancer Analysis

1. Data Cleaning

1.1 Data cleaning for binary data

We notice that in our cancer data some columns are binary data, however it include other numeric.

For example: In column"MedConditions_Diabets", it stands for question"healthy history_Diabetes or high blood sugar?"

Responsors should reply

- 1=yes
- 2=no

but we find

- -9=missing data

We have multipule binary columns which include missing values. For example:

- MedConditions_Diabetes
- MedConditions_HighBP
- MedConditions_HeartCondition
- MedConditions_LungDisease
- MedConditions_Arthritis
- MedConditions_Depression
- Smoke100
- GenderC
- FamilyEverHadCancer
- EverHadCancer
- BornInUSA

The clean process for binary data, we keep value=1 and 2, delect value= -9.

Value	Value Label	Action
1	Yes	Keep
2	No	Keep
-9	Missing data	Delet

Before we delect, we use mysql query to check how many lines with MedConditions_Diabetes value= -9, from result we get number is 78 lines with value = -9

In [14]:
```
%%sql
SELECT MedConditions_Diabetes, count(*)
FROM cancer_info
WHERE MedConditions_Diabetes is not null
GROUP BY MedConditions_Diabetes;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
3 rows affected.
```

Out[14]:

MedConditions_Diabetes	count(*)
2	2546
1	661
-9	78

Before we operate, we use mysql query to check total lines in table "cancer_info".

In [15]:
```
%%sql
SELECT COUNT(PersonID)
FROM cancer_info;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
1 rows affected.
```

Out[15]:

COUNT(PersonID)
3285

Use query delect value=-9 in 11 columns,list as follows

- MedConditions_Diabetes
- MedConditions_HighBP
- MedConditions_HeartCondition
- MedConditions_LungDisease
- MedConditions_Arthritis
- MedConditions_Depression
- Smoke100
- GenderC
- FamilyEverHadCancer
- EverHadCancer
- BornInUSA

In [16]:
```
%%sql
DELETE FROM cancer_info
WHERE (
        MedConditions_Diabetes = -9
        OR MedConditions_HighBP = -9
        OR MedConditions_HeartCondition = -9
        OR MedConditions_LungDisease = -9
        OR MedConditions_Arthritis = -9
        OR MedConditions_Depression = -9
        OR Smoke100 = -9
        OR GenderC = -9
        OR FamilyEverHadCancer = -9
        OR EverHadCancer = -9
        OR BornInUSA = -9
      );
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
277 rows affected.
```

Out[16]: []

In [17]:
```
%%sql
SELECT COUNT(PersonID)
FROM cancer_info;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
1 rows affected.
```

Out[17]:

COUNT(PersonID)
3008

1.2 Data cleaning for columns with identified number

We notice that in our cancer data some columns have identified number. They present range of selection.However in dataset it includes other numeric which is not in the selection.

For example: In column"Fruit", it stands for question"About how many cups of fruit (including 100% pure fruit juice) do you eat or drink each day?"

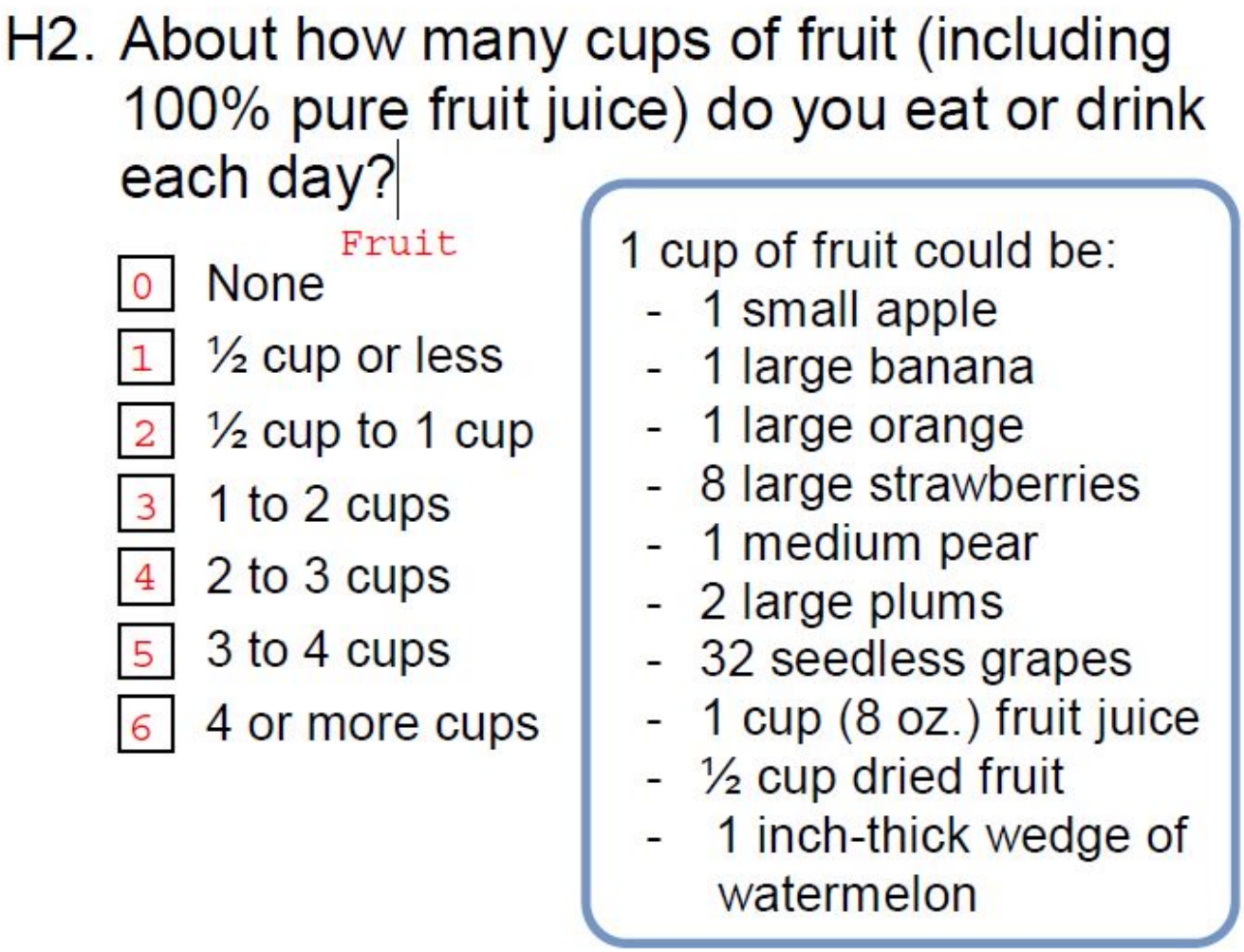
Responsors should reply

- 1= ½ cup or less
- 2= ½ cup to 1 cup
- 3= 1 to 2 cups
- 4= 2 to 3 cups
- 5= 3 to 4 cups
- 6= 4 or more cups

but we find

- -9=missing data
- -5=Multiple responses selected in error

Figure 10: Question H2



Use query delect value=-9 and value= -5 in 7 columns,list as follows

- Fruit
- Vegetables
- TimesModerateExercise
- TanningBed
- SkinCancerHPExam
- SmokeNow
- IncomeRanges
- GeneralHealth
- Age

The clean process for binary data, we keep other value, delect value= -9 and -5.

Value	Value Label	Action
1	½ cup or less	Keep
2	½ cup to 1 cup	Keep
3	1 to 2 cups	Keep
4	2 to 3 cups	Keep
5	3 to 4 cups	Keep
6	4 or more	Keep
-9	Missing data	Delet
-5	Multiple responses selected in error	Delet

In [18]:
```
%%sql
DELETE FROM cancer_info
WHERE (
        Fruit IN (-9,-5)
        OR Vegetables IN (-9,-5)
        OR TimesModerateExercise IN (-9,-5)
        OR TanningBed IN (-9,-5)
        OR SkinCancerHPExam IN (-9,-5)
        OR SmokeNow IN (-9,-5)
        OR IncomeRanges IN (-9,-5)
        OR GeneralHealth IN (-9,-5)
        OR Age IN (-9,-5)
        OR Weight IN (-9,-4)
        OR FamilyEverHadCancer IN (-9,-5)
      );
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
374 rows affected.
```

Out[18]: []

2. Create Column named Height

In our dataset, we have two columns to record responsors heights,we want to convert into one column "Height":

Table X: ER Diagram Part IV

Column Name	Description	Question
Height_Feet	height in feet	About how tall are you without shoes? Feet:
Height_Inches	height in inches	About how tall are you without shoes? Inches:

New Column Name	
Height	=Height_Feet*12+Height_Inches

In [19]:
```
%%sql
SELECT Height_Feet, Height_Inches,
    (Height_Feet*12+Height_Inches)AS Height
FROM cancer_info LIMIT 10;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
10 rows affected.
```

Out[19]:

Height_Feet	Height_Inches	Height
5	9	69
5	9	69
5	6	66
6	2	74
5	9	69
5	2	62
5	4	64
5	6	66
5	6	66
5	9	69

3. Risk Factor Analysis

3.1 Physical Activity

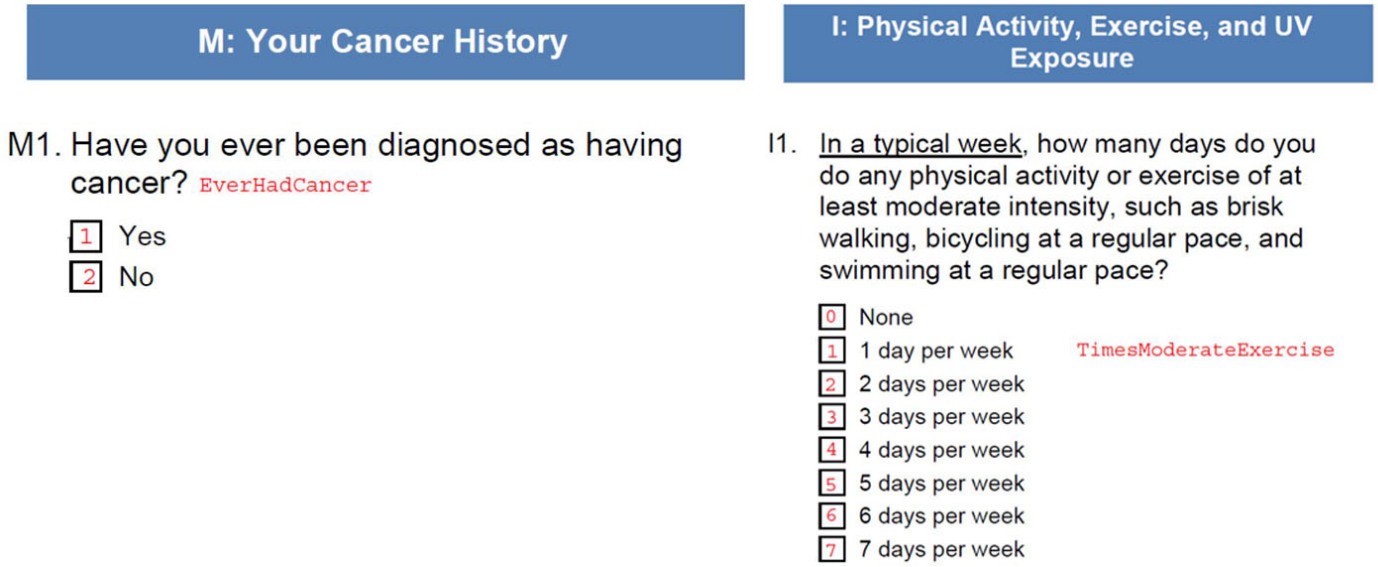
Question: Does physical activity or exercise of at least moderate intensity reduce the posibility of getting cancer? The Answer is yes.

3.1.1 Selected Variables

- **"EverHadCancer"** is answer for "Have you ever been diagnosed as having cancer?"
- **"TimesModerateExercise"** is answer for "how many days do you do any physical activity or exercise of at least moderate intensity, such as brisk walking, bicycling at a regular pace, and swimming at a regular pace?"

We want to present the relation between physical exercise with caner.

Figure 11: Questions for 'EverHadCancer' & 'TimesMederateExercise'



From query result, we could find the percentage who doesn't excise in who had cancer is significant higher than the percentage in who ever had cancer.

We use query to calculate physical activity amount percentage by "have Cancer" and "Not Have Cancer".

In [20]:
```
%%sql
SELECT TimesModerateExercise,
    COUNT(*) AS Total ,
    (COUNT(*) / (SELECT COUNT(*) FROM cancer_info WHERE EverHadCancer=
1)) * 100 AS 'Percentage to have cancer'
FROM cancer_info
WHERE EverHadCancer=1
GROUP BY TimesModerateExercise
ORDER BY TimesModerateExercise
;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
8 rows affected.
```

Out[20]:

TimesModerateExercise	Total	Percentage to have cancer
0	120	31.3316
1	31	8.0940
2	45	11.7493
3	65	16.9713
4	28	7.3107
5	45	11.7493
6	19	4.9608
7	30	7.8329

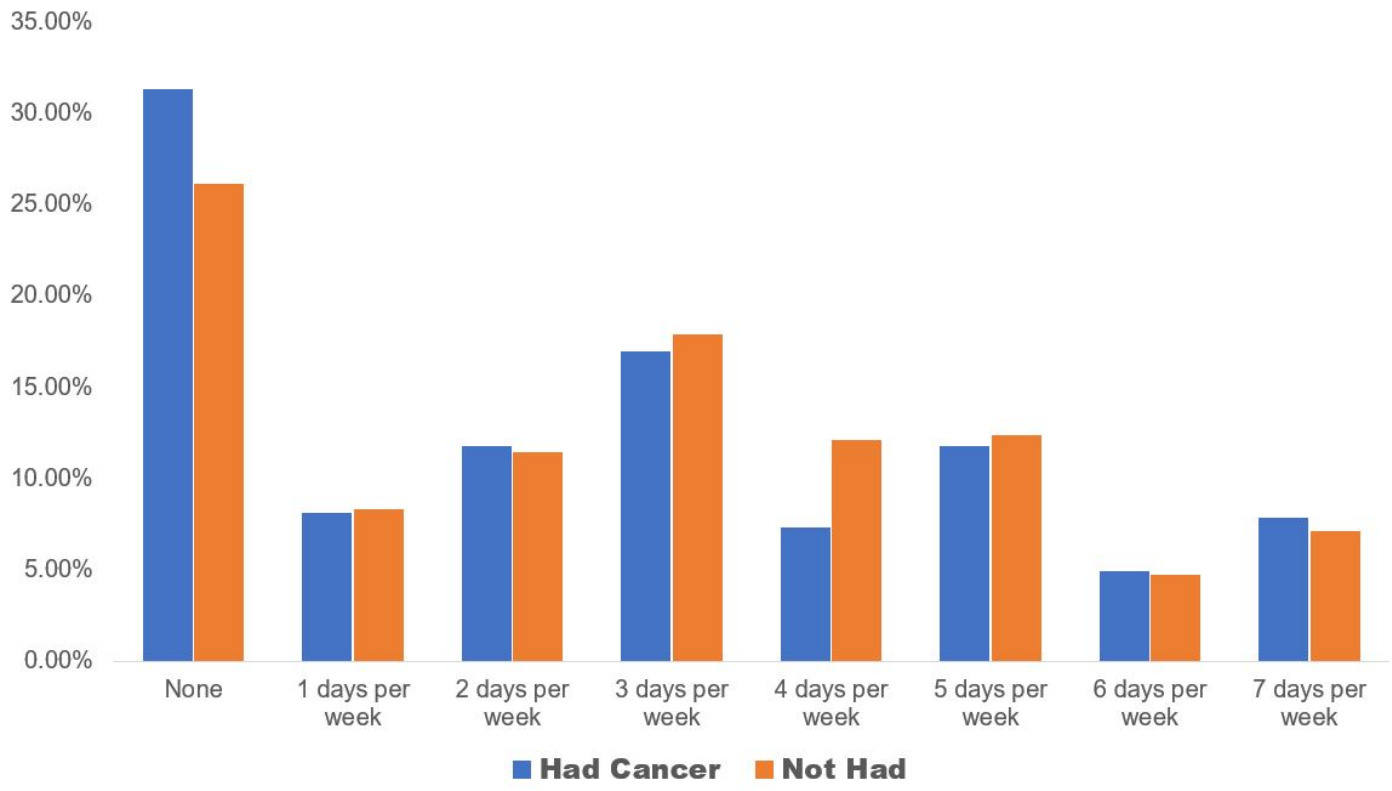
In [21]:
```
%%sql
SELECT TimesModerateExercise,
    COUNT(*) AS Total ,
    (COUNT(*) / (SELECT COUNT(*) FROM cancer_info WHERE EverHadCancer=
2)) * 100 AS 'Percentage to not have cancer'
FROM cancer_info
WHERE EverHadCancer=2
GROUP BY TimesModerateExercise
ORDER BY TimesModerateExercise
;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
8 rows affected.
```

Out[21]:

TimesModerateExercise	Total	Percentage to not have cancer
0	588	26.1217
1	187	8.3074
2	258	11.4616
3	402	17.8587
4	272	12.0835
5	278	12.3501
6	106	4.7090
7	160	7.1080

Figure 12: Excercise VS Cancer



3.2 Medical History

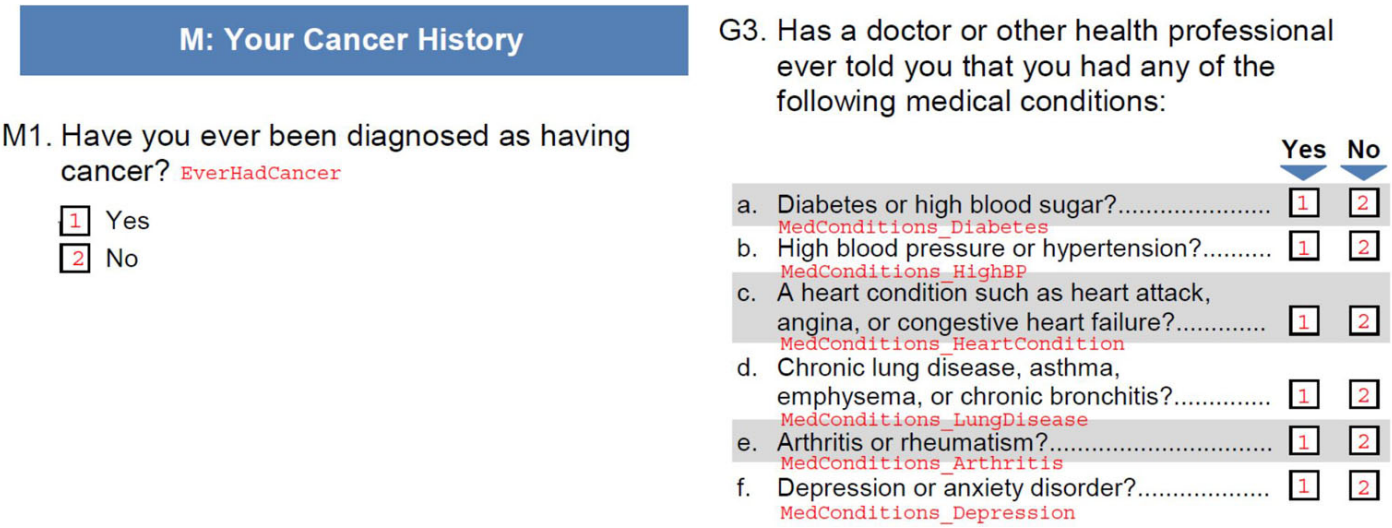
Question: Did medical history have kind of relation with the posibility of cancer? The Answer is No.

3.2.1 Selected Variables

- **MedConditions_Diabetes** is answer for "healthy history_Diabetes or high blood sugar?"
- **MedConditions_HighBP** is answer for "healthy history_High blood pressure or hypertension?"
- **MedConditions_HeartCondition** is answer for "healthy history_A heart condition such as heart attack, angina, or congestive heart failure?"
- **MedConditions_LungDisease** is answer for "healthy history_Chronic lung disease, asthma, emphysema, or chronic bronchitis?"
- **MedConditions_Arthritis** is answer for "healthy history_Arthritis or rheumatism?"
- **MedConditions_Depression** is answer for "healthy history_Depression or anxiety disorder?"

We want to present the relation between medical history with caner.

Figure 13: Question Related



Medical history of Diabetes with EverHadCancer relationship:

In [22]:
```
%%sql
SELECT EverHadCancer, MedConditions_Diabetes AS Diabetes,COUNT(*)
FROM cancer_info
GROUP BY EverHadCancer,Diabetes
ORDER BY EverHadCancer;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
4 rows affected.
```

Out[22]:

EverHadCancer	Diabetes	COUNT(*)
1	1	94
1	2	289
2	1	417
2	2	1834

Medical history of High BP with EverHadCancer relationship:

In [23]:
```
%%sql
SELECT EverHadCancer, MedConditions_HighBP AS HighBP,COUNT(*)
FROM cancer_info
GROUP BY EverHadCancer,HighBP
ORDER BY EverHadCancer;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
4 rows affected.
```

Out[23]:

EverHadCancer	HighBP	COUNT(*)
1	1	217
1	2	166
2	1	955
2	2	1296

The relationship between the Medical history of HeartCondition and EverHadCancer:

In [24]:
```
%%sql
SELECT EverHadCancer, MedConditions_HeartCondition AS HeartCondition,COU
NT(*)
FROM cancer_info
GROUP BY EverHadCancer,HeartCondition
ORDER BY EverHadCancer;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
4 rows affected.
```

Out[24]:

EverHadCancer	HeartCondition	COUNT(*)
1	1	55
1	2	328
2	1	196
2	2	2055

Medical history of LungDisease with EverHadCancer relationship:

In [25]:
```
%%sql
SELECT EverHadCancer, MedConditions_LungDisease AS Lung,COUNT(*)
FROM cancer_info
GROUP BY EverHadCancer,Lung
ORDER BY EverHadCancer;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
4 rows affected.
```

Out[25]:

EverHadCancer	Lung	COUNT(*)
1	1	76
1	2	307
2	1	277
2	2	1974

Medical history of Rheumatism with EverHadCancer relationship:

In [26]:
```
%%sql
SELECT EverHadCancer, MedConditions_Arthritis AS Rheumatism,COUNT(*)
FROM cancer_info
GROUP BY EverHadCancer,Rheumatism
ORDER BY EverHadCancer;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
4 rows affected.
```

Out[26]:

EverHadCancer	Rheumatism	COUNT(*)
1	1	165
1	2	218
2	1	602
2	2	1649

Medical history of Depression with EverHadCancer relationship:

In [27]:
```
%%sql
SELECT EverHadCancer, MedConditions_Depression AS Depression,COUNT(*)
FROM cancer_info
GROUP BY EverHadCancer,Depression
ORDER BY EverHadCancer;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
4 rows affected.
```

Out[27]:

EverHadCancer	Depression	COUNT(*)
1	1	103
1	2	280
2	1	514
2	2	1737

3.2.2 Results

We don't find any relation between history disease with cancer. Because publics who had cancer don't have possitive or negative relation with their disease backgroup.

Table 9: Cancer VS Condition

Condition	Had Cancer	Not Had Cancer	Condition	Had Cancer	Not Had Cancer
Had Diabetes	94	417	Not Had Diabetes	289	1834
Had HighBP	217	956	Not Had HighBP	166	1296
Had HeardCon	55	196	Not Had HeardCon	328	2056
Had Lung	76	277	Not Had Lung	307	1974
Had Rheumatism	165	602	Not Had Rheumatism	218	1649
Had Depression	165	602	Not Had Depression	280	1737

3.3 Family Cancer History

Question: Does Family Cancer History affects individual Cancer Risk? The Answer is Yes.

3.3.1 Variable Selection

- Cancer is a common disease, so it's no surprise that many families have at least a few members who have had cancer.
- Sometimes, certain types of cancer seem to run in some families. In some cases, this might be because family members share certain behaviors or exposures that increase cancer risk, such as such as smoking.
- Selected Variables:
  - **"EverHadCancer"** is answer for "Have you ever been diagnosed as having cancer?"
  - **"FamilyEverHadCancer"** is answer for "Have any of your family members ever had cancer?"
- We want to present the relation between family cancer history with individual caner risk.

Figure 14: Question Related

M: Your Cancer History

M1. Have you ever been diagnosed as having cancer? EverHadCancer

- [1] Yes
- [2] No

N5. Have any of your family members ever had cancer? FamilyEverHadCancer

- [1] Yes
- [2] No
- [4] Not sure

In [28]:
```
%%sql
SELECT FamilyEverHadCancer,
    COUNT(*) AS Total ,
    (COUNT(*) / (SELECT COUNT(*) FROM cancer_info WHERE EverHadCancer=
1)) * 100 AS 'Percentage to have cancer'
FROM cancer_info
WHERE EverHadCancer=1
GROUP BY FamilyEverHadCancer
ORDER BY FamilyEverHadCancer;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
3 rows affected.
```

Out[28]:

FamilyEverHadCancer	Total	Percentage to have cancer
1	303	79.1123
2	60	15.6658
4	20	5.2219

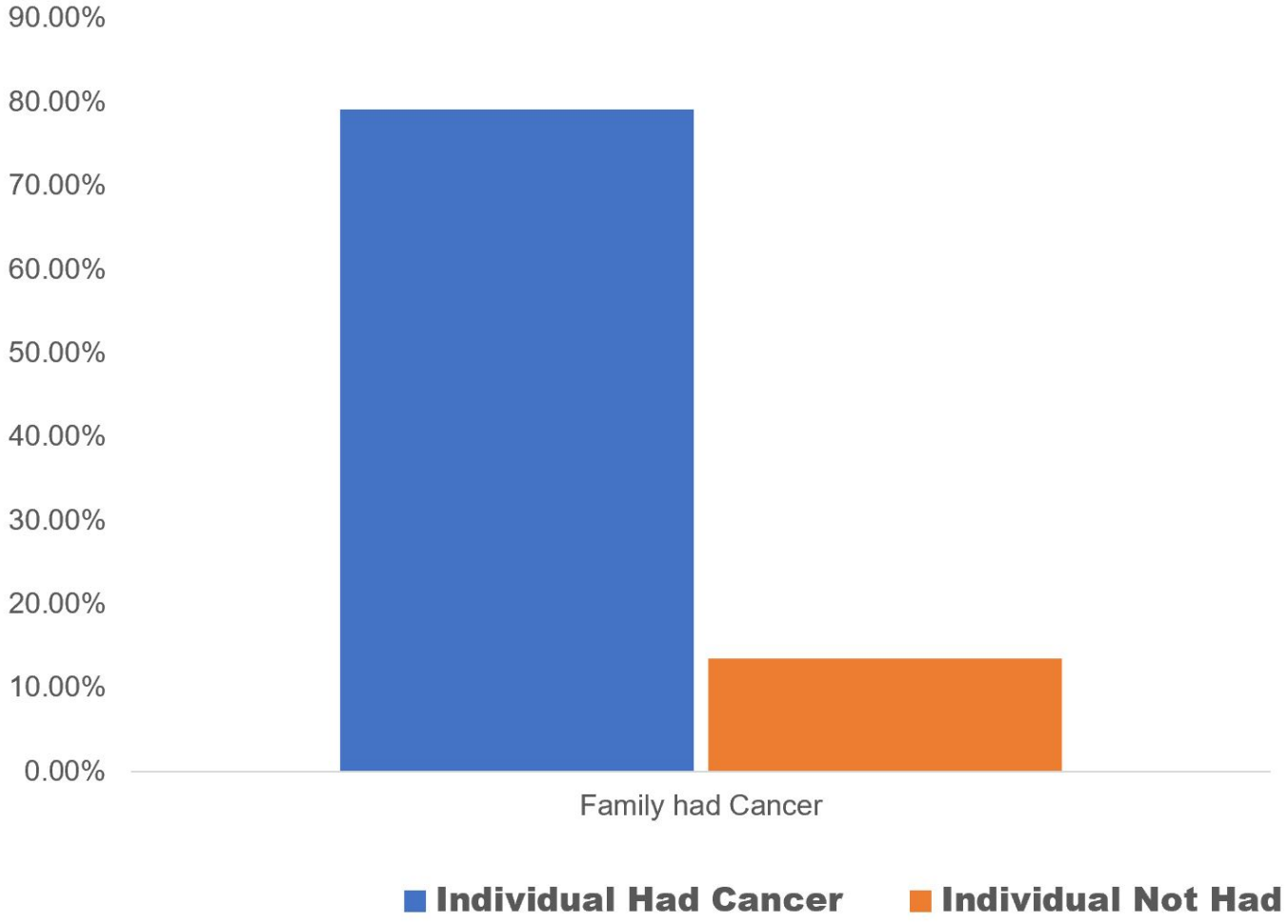
In [29]:
```
%%sql
SELECT FamilyEverHadCancer,
    COUNT(*) AS Total ,
    (COUNT(*) / (SELECT COUNT(*) FROM cancer_info WHERE EverHadCancer=
2)) * 100 AS 'Percentage to not have cancer'
FROM cancer_info
WHERE EverHadCancer=1
GROUP BY FamilyEverHadCancer
ORDER BY FamilyEverHadCancer;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
3 rows affected.
```

Out[29]:

FamilyEverHadCancer	Total	Percentage to not have cancer
1	303	13.4607
2	60	2.6655
4	20	0.8885

Figure 15: Cancer VS Family History



Prostate Cancer Analysis

1. Export prostate_analysis into CSV file

For the convinence of doing analysis in other applications, e.g. Tableau, TABLE prostate_analysis was exported.

In [ ]:
```
%%sql

SELECT
    'PersonID',
    'EverHadCancer',
    'CaProstate',
    'Cancer_Cat',
    'WhenDiagnosedCancer',
    'UndergoCancerTreatment',
    'CancerTx_Chemo',
    'CancerTx_Radiation',
    'CancerTx_Surgery',
    'CancerTx_Other',
    'HowLongFinishTreatment_Cat',
    'CancerTxSummary',
    'CancerDeniedCoverage',
    'CancerHurtFinances',
    'CancerAbilityToWork',
    'ClinicalTrialCancerTx',
    'DiscussedClinicalTrial',
    'GenderC',
    'EverHadPSATest',
    'FamilyEverHadCancer',
    'Age',
    'Race_Cat2',
    'Height_Feet',
    'Height_Inches',
    'Height',
    'Weight',
    'BMI',
    'UseMenuCalorieInfo',
    'Fruit',
    'Vegetables',
    'Smoke100',
    'SmokeNow',
    'TriedQuit',
    'ConsiderQuit',
    'Cancer_Cat_Label',
    'Race_Cat_Label'
FROM prostate_analysis
UNION
SELECT
    PersonID,
    EverHadCancer,
    CaProstate,
    Cancer_Cat,
    WhenDiagnosedCancer,
    UndergoCancerTreatment,
    CancerTx_Chemo,
    CancerTx_Radiation,
    CancerTx_Surgery,
    CancerTx_Other,
    HowLongFinishTreatment_Cat,
    CancerTxSummary,
    CancerDeniedCoverage,
    CancerHurtFinances,
```

```
    CancerAbilityToWork,
    ClinicalTrialCancerTx,
    DiscussedClinicalTrial,
    GenderC,
    EverHadPSATest,
    FamilyEverHadCancer,
    Age,
    Race_Cat2,
    Height_Feet,
    Height_Inches,
    Height,
    Weight,
    BMI,
    UseMenuCalorieInfo,
    Fruit,
    Vegetables,
    Smoke100,
    SmokeNow,
    TriedQuit,
    ConsiderQuit,
    Cancer_Cat_Label,
    Race_Cat_Label
FROM prostate_analysis
INTO OUTFILE '/home/data/prostate_analysis_2.csv'
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n';
```

2. Data Cleaning

We did not clean the data for 'prostate_analysis' table, because the sample size of the people with prostate cancer is rather small, and we may lose a large portion of data if we remove missing value or error terms. In replacement, the missing data and error terms would be filtered out in each seperate questions.

3. Risk Factor Analysis

3.1 Age

As indicated by the rates of diagnosis, age is the biggest—but not the only—risk factor for prostate cancer.

Question: For those with prostate cancers, what is the age range and distribution?

In [30]:
```
%%sql
SELECT
    #EverHadCancer,
    Cancer_Cat_Label,
    COUNT(*) AS count,
    AVG(Age) AS avg_age,
    AVG(WhenDiagnosedCancer) AS avg_diagnosedAge

FROM prostate_analysis
WHERE (
    (Cancer_Cat = -1 OR CaProstate  = 1) AND
    AGE > 0
)
GROUP BY Cancer_Cat_Label;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
3 rows affected.
```

Out[30]:

Cancer_Cat_Label	count	avg_age	avg_diagnosedAge
Inapplicable, coded 2 in EverHadCancer	2644	54.3767	-1.0004
More than one cancer checked	12	73.2500	60.9167
Prostate cancer only	40	70.3250	62.7000

We could observe from the following figures that, the healthy population is crowded between 40 to 70, while the people with prostate cancer is crowded between 65 to 80. The diagnosed age is slightly younger than the current age.

Figure 16: Age Distribution for 3 Groups

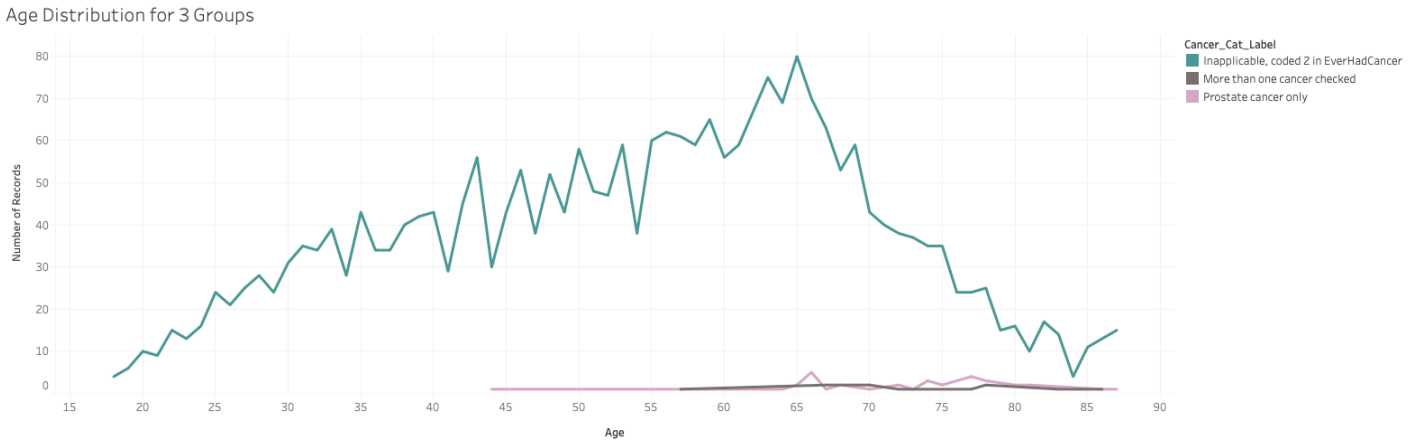
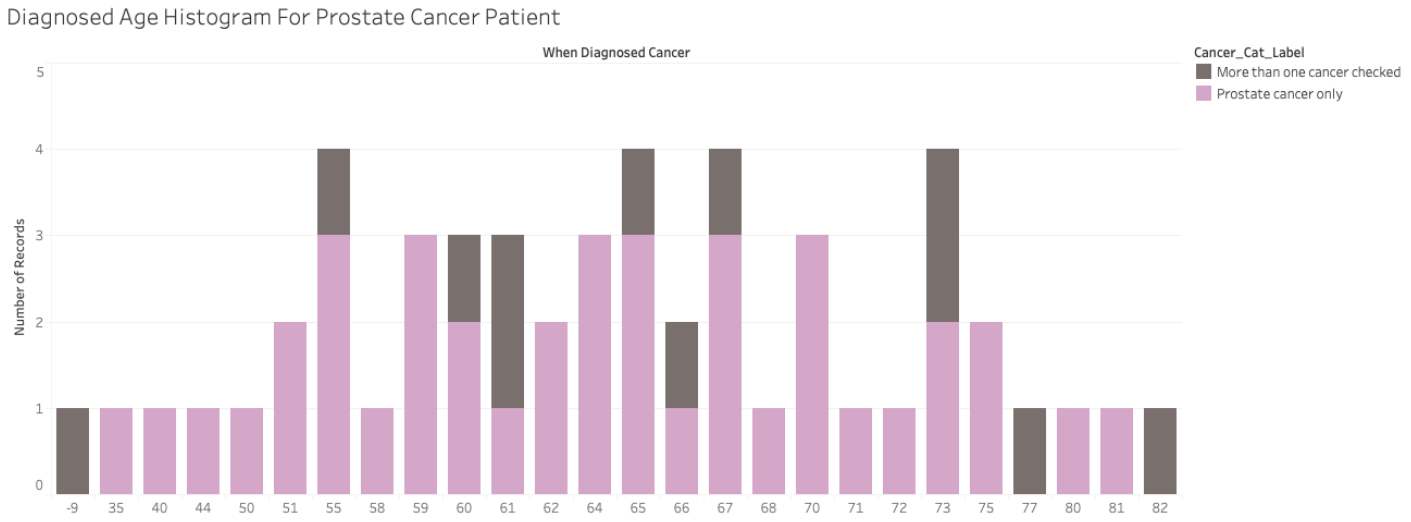


Figure 17: Diagnosed Age Histogram For Prostate Cancer Patient



Question: Are there any other type of Cancers whose patients are younger on average?

In [31]:
```
%%sql
SELECT
    #EverHadCancer,
    Cancer_Cat_Label,
    COUNT(*) AS count,
    AVG(Age) AS avg_age,
    AVG(WhenDiagnosedCancer) AS avg_diagnosedAge
FROM prostate_analysis
WHERE (
    (Cancer_Cat >= -1) AND
    AGE > 0
)
GROUP BY Cancer_Cat_Label
ORDER BY avg_diagnosedAge;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
23 rows affected.
```

Out[31]:

Cancer_Cat_Label	count	avg_age	avg_diagnosedAge
Inapplicable, coded 2 in EverHadCancer	2644	54.3767	-1.0004
Stomach cancer only	2	82.0000	20.0000
Cervical cancer only	28	59.3571	26.3214
Hodgkins	4	43.2500	31.0000
Non-Hodgkin	8	57.6250	42.6250
Endometrial cancer only	9	66.0000	44.2222
Ovarian cancer only	2	74.5000	44.5000
Liver cancer only	1	46.0000	46.0000
Other cancer only	34	63.2941	46.8235
More than one cancer checked	68	70.0735	46.9412
Leukemia	7	61.7143	47.1429
Melanoma	21	63.1429	49.2381
Bone cancer only	1	62.0000	50.0000
Skin cancer only	120	67.6500	53.2667
Colon cancer only	22	68.8182	54.0455
Breast cancer only	81	66.9383	54.5432
Renal cancer only	3	62.0000	55.0000
Head/Neck cancer only	6	61.1667	59.8333
Rectal cancer only	4	68.5000	62.5000
Prostate cancer only	40	70.3250	62.7000
Bladder cancer only	5	69.0000	63.2000
Pancreatic cancer only	2	67.5000	64.5000
Lung cancer only	8	74.6250	73.0000

3.1.2 Result on AGE

Through observation, we noticed that the average age/diagnosed age of the Prostate Cancer Patient is larger than the average age of the Non-Cancer People.

3.2 Family History & Genetic Factors

Genes for disease can run in families. Men who have a relative with prostate cancer are twice as likely to develop the disease, while those with 2 or more relatives are nearly 4 times as likely to be diagnosed. The risk is even higher if the affected family members were diagnosed before age 65.

3.1.1 Variable Selection

- Variable: FamilyEverHadCancer

N5. Have any of your family members ever had cancer? `FamilyEverHadCancer`

- [1] Yes
- [2] No
- [4] Not sure

In [32]:
```
%%sql
SELECT
        CaProstate,
        CASE
            WHEN FamilyEverHadCancer = 1 THEN "Yes"
            WHEN FamilyEverHadCancer = 2 THEN "No"
            WHEN FamilyEverHadCancer = 4 THEN "Not sure"
        END AS FamilyCancerIndex,
        COUNT(*) AS count
        #Cancer_Cat_Label
FROM prostate_analysis
WHERE(
    (Cancer_Cat = -1 OR CaProstate  = 1) AND
    (prostate_analysis.FamilyEverHadCancer > 0))
GROUP BY CaProstate, FamilyCancerIndex
ORDER BY CaProstate, FamilyCancerIndex;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
6 rows affected.
```

Out[32]:

CaProstate	FamilyCancerIndex	count
-1	No	676
-1	Not sure	140
-1	Yes	1856
1	No	11
1	Not sure	3
1	Yes	37

In [34]:
```
%%sql
SELECT
    Cancer_Cat_Label,
    SUM(CASE WHEN FamilyEverHadCancer=1 THEN 1 ELSE 0 END) AS FamilyEver
HadCancer_Y,
    SUM(CASE WHEN FamilyEverHadCancer=2 THEN 1 ELSE 0 END) AS FamilyEver
HadCancer_N,
    SUM(CASE WHEN FamilyEverHadCancer=4 THEN 1 ELSE 0 END) AS FamilyEver
HadCancer_NotSure
FROM prostate_analysis
WHERE(
        (Cancer_Cat = -1 OR CaProstate  = 1) AND
        (prostate_analysis.FamilyEverHadCancer > 0))
GROUP BY Cancer_Cat_Label;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
3 rows affected.
```

Out[34]:

Cancer_Cat_Label	FamilyEverHadCancer_Y	FamilyEverHadCancer_N	FamilyEverHadCa
Inapplicable, coded 2 in EverHadCancer	1856	676	140
More than one cancer checked	7	4	1
Prostate cancer only	30	7	2

3.3.2 Result on Family History & Genetic Factors

Visualizing the query results in Tableau, we could see that people with prostate cancer have a higher percentage in terms of family cancer history, when comparing 76% to 69%. Figure Z also shows that the percentage of having cancer in family history is higher for breast and skin cancer.

Figure 18: Comparison of Family Cancer History for 3 Groups

Comparison of Family Cancer History for 3 Groups

Cancer_Cat_Label	Family Ever Had Cancer 1	2	4
Inapplicable, coded 2 in EverHadCancer	69.46% 1,856	25.30% 676	5.24% 140
More than one cancer checked	58.33% 7	33.33% 4	8.33% 1
Prostate cancer only	76.92% 30	17.95% 7	5.13% 2

% of Total Number of R.. 5.13% 76.92%

Figure 19: Comparison of Family Cancer History for Several Cancer Groups

Comparison of Family Cancer History for Several Cancer Groups

Cancer_Cat_Label	Family Ever Had Cancer 1	2	4
Breast cancer only	82.93% 68	10.98% 9	6.10% 5
Cervical cancer only	55.56% 15	25.93% 7	18.52% 5
Colon cancer only	69.57% 16	26.09% 6	4.35% 1
Inapplicable, coded 2 in EverHadCancer	69.46% 1,856	25.30% 676	5.24% 140
Lung cancer only	75.00% 6	12.50% 1	12.50% 1
Melanoma	71.43% 15	19.05% 4	9.52% 2
More than one cancer checked	74.63% 50	23.88% 16	1.49% 1
Non-Hodgkin	50.00% 4	37.50% 3	12.50% 1
Other cancer only	90.91% 30	6.06% 2	3.03% 1
Prostate cancer only	76.92% 30	17.95% 7	5.13% 2
Skin cancer only	86.78% 105	9.92% 12	3.31% 4

% of Total Number of R.. 1.49% 90.91%

3.3 Race/ Financial States

Men of African descent are **76% more likely** to develop prostate cancer compared with white men, and **2.2 times more likely** to die from the disease.

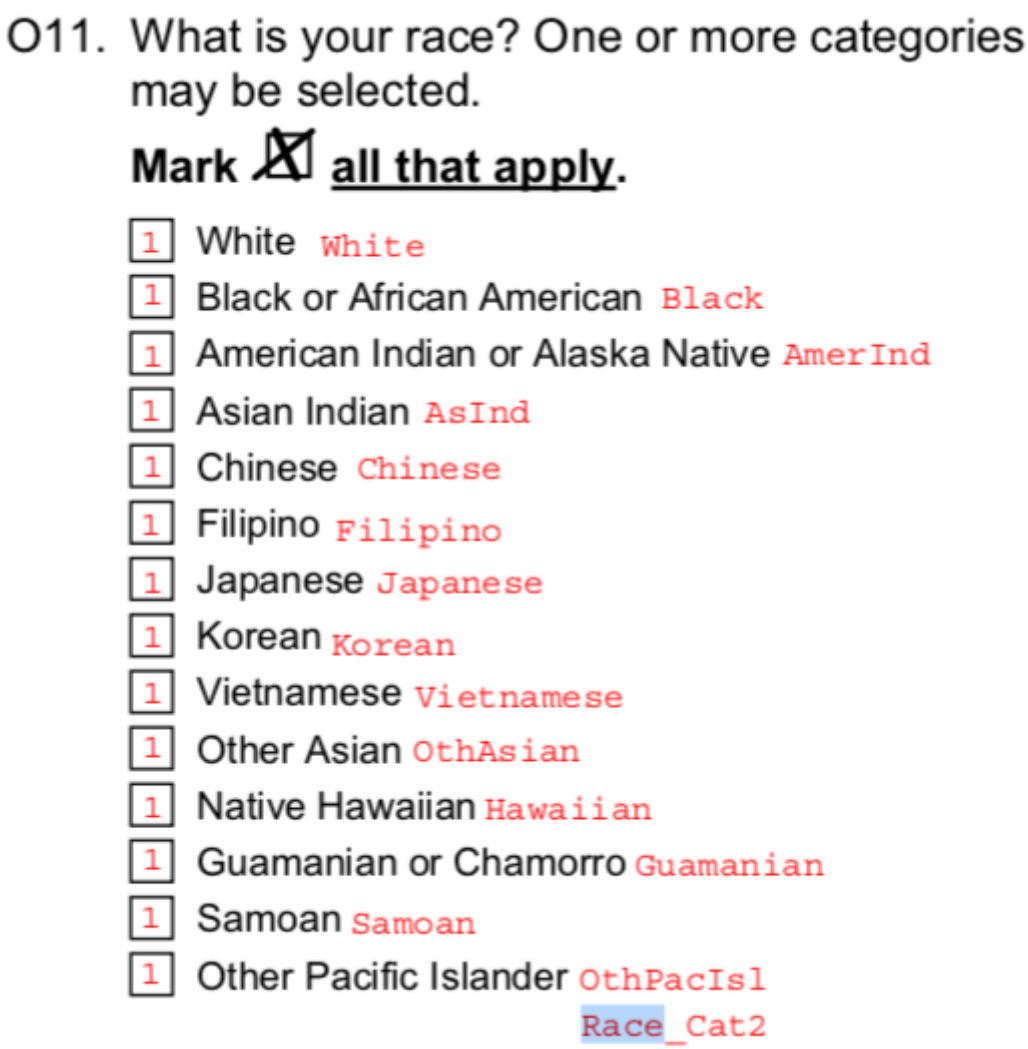
The **increased death rate** from prostate cancer has been shown to be due in part to:

```
* inequality in access to healthcare,
* insurance, PSA screening,
* appropriate treatment and follow-up,
* other simultaneous conditions or treatments, and
* other socioeconomic factors.
```

3.3.1 Variable Selection for Race

We used **Variable: Race_Cat2**, since it is a derived categorical variable with multiple categories.

Figure 20: Question for 'Race_Cat2'



In [36]:
```
%%sql
SELECT
    Race_Cat_Label,
    COUNT(*) AS count,
    ( COUNT(*)*100/ (SELECT COUNT(*) FROM prostate_analysis WHERE(CaPros
tate = 2 AND Race_Cat2 IN (11, 12, 16, 31)))) AS '%'

FROM prostate_analysis
WHERE(CaProstate = 2 AND Race_Cat2 IN (11, 12, 16, 31) )
GROUP BY Race_Cat_Label;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
3 rows affected.
```

Out[36]:

Race_Cat_Label	count	%
White	337	83.2099
Black	50	12.3457
Multiple races selected	18	4.4444

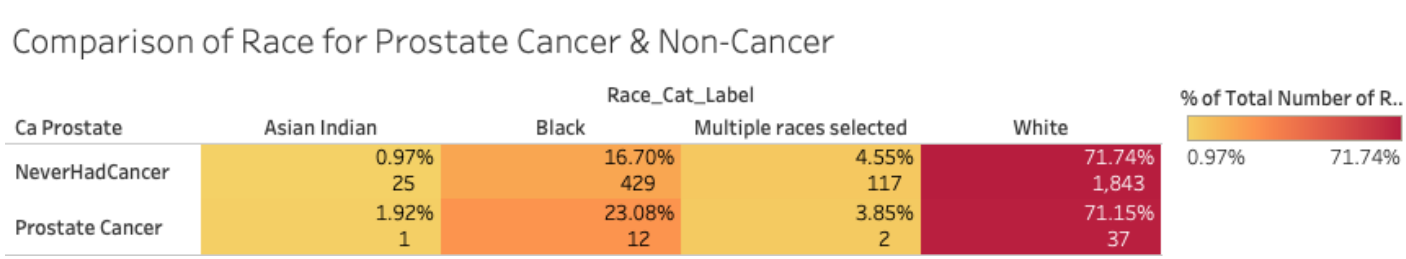
In [38]:
```
%%sql
SELECT
    Race_Cat_Label,
    COUNT(*) AS count,
    (COUNT(*)*100/ (SELECT COUNT(*) FROM prostate_analysis WHERE(CaProst
ate = 1 AND Race_Cat2 >0))) AS '%'

FROM prostate_analysis
WHERE(CaProstate = 1 AND Race_Cat2 >0)
GROUP BY Race_Cat_Label;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
4 rows affected.
```

Out[38]:

Race_Cat_Label	count	%
White	37	71.1538
Black	12	23.0769
Asian Indian	1	1.9231
Multiple races selected	2	3.8462

Figure 21: Comparison of Race for Prostate Cancer & Non-Cancer



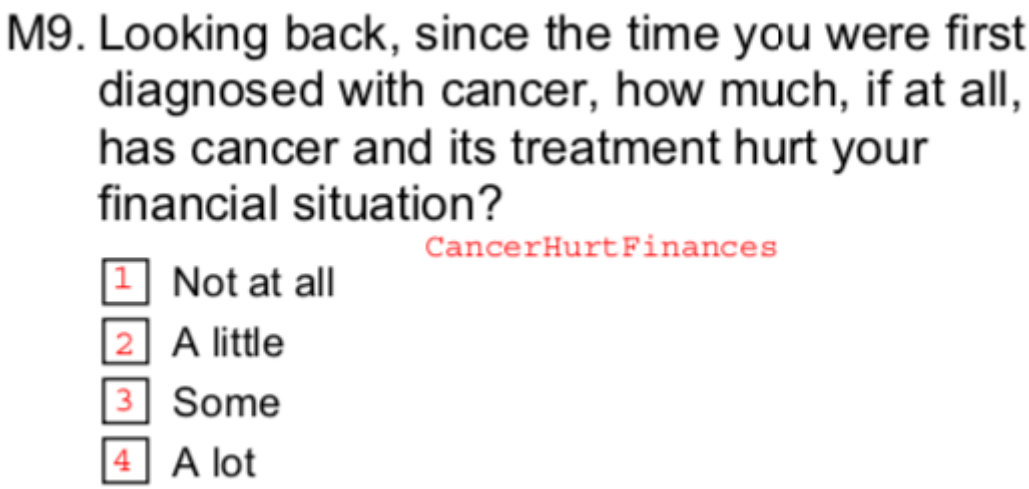
3.3.2 Result on Race

Both the figure and the query result show that there is a higher percentage of Black or African Americans in prostate cancer patient than in people never had cancer.

3.3.3 Variable Selection for Financial States

There is an interesting question in the survey that we think could be used to evaluate the relationship between cancer and one's financial states: **Variable: CancerHurtFinance**

Figure 22: Question for 'CancerHurtFinance'



In [39]:
```
%%sql
SELECT
    Race_Cat_Label,
    AVG(CancerHurtFinances/4) as average_rate_CancerHurtFinances,
    COUNT(*)
FROM prostate_analysis
WHERE(CaProstate = 1 AND Race_Cat2 >0 AND CancerHurtFinances >0)
GROUP BY Race_Cat_Label
ORDER BY Race_Cat_Label;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
3 rows affected.
```

Out[39]:

Race_Cat_Label	average_rate_CancerHurtFinances	COUNT(*)
Asian Indian	0.25000000	1
Black	0.55555556	9
White	0.39393939	33

3.3.2 Result on Race

We used the indexes in the survey directly, and calculated an average rate for each race. The result shows that cancer probably have hurt the financial situation of the Black people the most.

3.4 Lifestyle & Dietary habits

3.4.1 Weight Factor

Men who are overweight or obese are at greater risk of ultimately developing an aggressive form of prostate cancer. Research has shown that in obese men, recovery from surgery tends to be longer and more difficult, and the risk of dying from prostate cancer can be higher.

There is an definition of overweigh and obese that is measured by BMI:

- If your BMI is less than 18.5, it falls within the underweight range.
- If your BMI is 18.5 to <25, it falls within the normal.
- If your BMI is 25.0 to <30, it falls within the overweight range.
- If your BMI is 30.0 or higher, it falls within the obese range.

3.4.2 Variable Selection for Weight Factor

We choose to use BMI as a major variable for this analysis. However, this index must be calculated based on the height and the weight of the person. Moreover, the height is not directely given in the dataset, and we need to calculate it like in previous analysis. The variables and the formulas we used are as follows:

- Variables: **Height_Feet, Height_Inches, Weight** are used to calculate **BMI** when importing data
- Formulas:

```
  (section_g.Height_Feet * 12 + section_g.Height_Inches) AS Height,
  (section_g.Weight*703)/ ( POWER((section_g.Height_Feet * 12 + sectio
n_g.Height_Inches) , 2) ) AS BMI
```

Question: What is the average BMI for people without cancer and people with prostate cancer?

In [40]:
```
%%sql
SELECT
    CaProstate,
    COUNT(*) AS Count,
    AVG(BMI) AS avg_BMI
FROM prostate_analysis
WHERE( Height_Feet > 0 AND Weight > 0 AND ( CaProstate= -1 OR CaProstate
= 1))
GROUP BY CaProstate;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
2 rows affected.
```

Out[40]:

CaProstate	Count	avg_BMI
-1	2667	28.445135328205623
1	54	28.765771241971635

Question: How about the average BMI of other cancer groups?

In [41]:
```
%%sql
SELECT
    Cancer_Cat_Label,
    COUNT(*) AS Count,
    AVG(BMI) AS avg_BMI
FROM prostate_analysis
WHERE( Height_Feet > 0 AND Weight > 0 AND Cancer_Cat >= -1)
GROUP BY Cancer_Cat_Label
ORDER BY avg_BMI DESC;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
23 rows affected.
```

Out[41]:

Cancer_Cat_Label	Count	avg_BMI
Liver cancer only	1	41.960514233241504
Leukemia	7	34.91057211566687
Rectal cancer only	4	33.97812985143052
Bone cancer only	1	33.06234883095942
Renal cancer only	4	31.543388489835298
Endometrial cancer only	10	30.70493629735776
Lung cancer only	8	29.69996887966954
Cervical cancer only	27	29.467527609096805
Hodgkins	4	29.412167423645897
Prostate cancer only	40	29.363218796509216
Breast cancer only	82	29.219711980306446
Non-Hodgkin	9	28.86993826157014
Head/Neck cancer only	6	28.815126866792678
Other cancer only	34	28.812863392602267
More than one cancer checked	69	28.662365081428153
Inapplicable, coded 2 in EverHadCancer	2654	28.422107131888882
Melanoma	23	27.913175564653244
Skin cancer only	121	27.465648423463772
Colon cancer only	23	26.368208229190728
Bladder cancer only	5	26.337666005636652
Ovarian cancer only	2	26.15337512644198
Stomach cancer only	2	24.5792655267659
Pancreatic cancer only	1	23.55731922398589

3.4.3 Result on Weight Factor

The BMI for the prostate cancer group and the healthy group are very close, showing that weight might not be a leading factor for prostate cancer. However, there are cancer groups with a BMI higher than 30, probably indicating that such cancer groups are affected by weight.

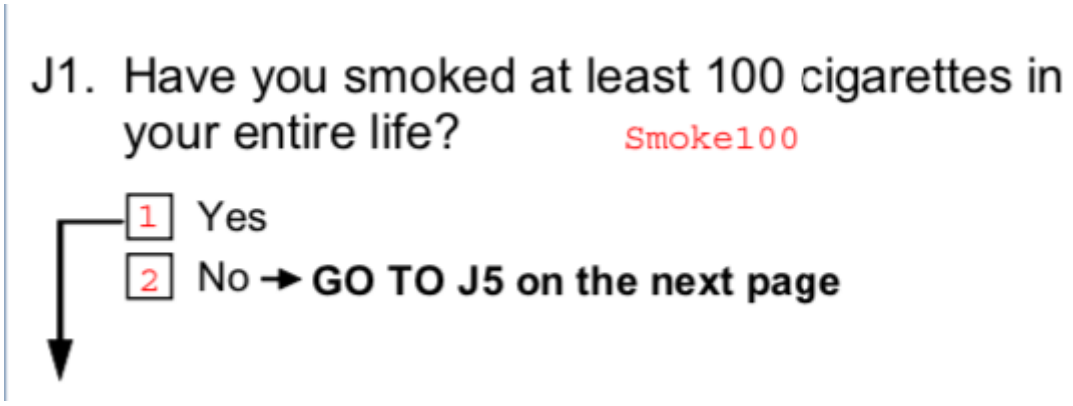
3.4.4 Smoking Factor

While smoking has not been thought to be a risk factor for low-risk prostate cancer, it may be a risk factor for aggressive prostate cancer. Likewise, lack of vegetables in the diet (especially broccoli-family vegetables) is linked to a higher risk of aggressive prostate cancer, but not to low-risk prostate cancer.

3.4.5 Variable Selection for Smoking Factor

We choose Variable: **Smoke100**, as positive answers to the question may be an strong indicator of a regular smoker.

Figure 22: Question for 'Smoke100'



In [42]:
```
%%sql
SELECT
    Cancer_Cat_Label,
    Smoke100,
    COUNT(*) AS count
#     (COUNT(*)/ (SELECT Cancer_Cat_Label, ))
FROM prostate_analysis
WHERE ( (Cancer_Cat = -1 OR CaProstate  = 1 )AND Smoke100 = 1 )
GROUP BY Cancer_Cat_Label;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
3 rows affected.
```

Out[42]:

Cancer_Cat_Label	Smoke100	count
Inapplicable, coded 2 in EverHadCancer	1	1040
Prostate cancer only	1	21
More than one cancer checked	1	8

In [49]:
```
%%sql
SELECT
    Cancer_Cat_Label,
    SUM(CASE WHEN Smoke100=1 THEN 1 ELSE 0 END) AS Smoke100_Y,
    SUM(CASE WHEN Smoke100=2 THEN 1 ELSE 0 END) AS Smoke100_N
FROM prostate_analysis
WHERE(
        (Cancer_Cat = -1 OR CaProstate  = 1) AND
        (prostate_analysis.Smoke100 > 0))
GROUP BY Cancer_Cat_Label;
```

```
 * mysql+pymysql://root:***@fe512_mysql/fe512db
3 rows affected.
```

Out[49]:

Cancer_Cat_Label	Smoke100_Y	Smoke100_N
Inapplicable, coded 2 in EverHadCancer	1040	1694
More than one cancer checked	8	6
Prostate cancer only	21	21

Figure 23: Comparison of Smook100 for Several Cancer Groups

Comparison of Smook100 for Several Cancer Groups

Cancer_Cat_Label	Smoke100 Yes	Smoke100 No
Breast cancer only	39.76% 33	60.24% 50
Cervical cancer only	51.72% 15	48.28% 14
Colon cancer only	43.48% 10	56.52% 13
Inapplicable, coded 2 in EverHadCancer	38.04% 1,040	61.96% 1,694
Lung cancer only	62.50% 5	37.50% 3
Melanoma	47.83% 11	52.17% 12
More than one cancer checked	38.57% 27	61.43% 43
Non-Hodgkin	55.56% 5	44.44% 4
Other cancer only	44.12% 15	55.88% 19
Prostate cancer only	50.00% 21	50.00% 21
Skin cancer only	46.77% 58	53.23% 66

% of Total Number of R..
37.50% 62.50%

3.4.6 Result on Smoking

We could observe from the query and the figure that the percentage of smokers are higher in prostate cancer patients than in people never had cancer. In the figure we also listed a few other cancer groups that have a higher percentage of frequent smoker. It seems that lung cancer is definately affected by smoking.

CONCLUSION

We could conclude from the above analysis that, for cancer in general, the family cancer history and the level of physical activity all seemed to be risky factors. However, medical history such as diabetes, high blood pressure, heart condition, lung disease, arthritis, and depression all seemed not to be risky for cancer.

For prostate cancer, a lot of factors seemed to be risky according to our analysis. The probability of getting prostate cancer may increase as one grows old; or if one's family has a cancer history; or if one is black; or if one smokes a lot. Weight did not seem to be risky as we supposed.

Limitation & Future Study

Although we used control and test groups to observe the effect of risk factors, our analysis is not enough to build casual relationships between cancer & its risk factors. Classification and other machine learning method could be used to exam and measure effects of the risk factors on different types of cancers.

What's more, we could also dig into the reasons why some of the factors we analyzed did not appeared to be risky. It is possible that our finding is limited by the sample size, or by our methodology, as we only used descriptive statistics to compare the effect.

Reference

Centers for Disease Control and Prevention. (2017). Defining Adult Overweight and Obesity. Retrieved from https://www.cdc.gov/obesity/adult/defining.html (https://www.cdc.gov/obesity/adult/defining.html)

Centers for Disease Control and Prevention. (2017). Calculating BMI Using the English System. Retrieved from https://www.cdc.gov/nccdphp/dnpao/growthcharts/training/bmiage/page5_2.html (https://www.cdc.gov/nccdphp/dnpao/growthcharts/training/bmiage/page5_2.html)

National Cancer Institution. (2018). Cancer Statistics. Retrieved from https://www.cancer.gov/about-cancer/understanding/statistics (https://www.cancer.gov/about-cancer/understanding/statistics)

National Cancer Institution. (2017). HINTS5 Cycle1 Annotated Instrument English. Retrieved from https://hints.cancer.gov/docs/Instruments/HINTS5_Cycle1_Annotated_Instrument_English.pdf (https://hints.cancer.gov/docs/Instruments/HINTS5_Cycle1_Annotated_Instrument_English.pdf)

National Cancer Institution. (2015). Risk Factors for Cancer. Retrieved from https://www.cancer.gov/about-cancer/causes-prevention/risk (https://www.cancer.gov/about-cancer/causes-prevention/risk)

Prostate Cancer Foundation. (2019). Prostate Cancer: What Are The Risk Factors? Retrieved from https://www.pcf.org/patient-resources/family-cancer-risk/prostate-cancer-risk-factors/ (https://www.pcf.org/patient-resources/family-cancer-risk/prostate-cancer-risk-factors/)

Durrant, M. (2015). Database design for a survey. Retrieved from https://stackoverflow.com/questions/1764435/database-design-for-a-survey (https://stackoverflow.com/questions/1764435/database-design-for-a-survey)

Appendix

Appendix 1

FE512_dataimport_a_to_g.ipynb

Appendix 2

FE512_dataimport_h_to_o.ipynb