#### **Description of Python Files Used in the Deliverable:**

The file "provess\_csv\_calculate\_scores.py" contains two functions: split\_csv() which was used to split the CHH\_ProvincialHeartHealthData CSV file into 11 separate csv files according to the dimensions defined in the project, and calcuLate\_risk\_create\_fact() which was used to calculate Framingham and ASCVD risk scores from the dimensional attributes and create the csv file for the fact table.

The file "calculator.py" contains the two functions responsible for calculating the Framingham and ASCVD risk scores. These two functions were obtained from two GitHub repositories and modified to work with the project's dimensional attributes and the calcuLate\_risk\_create\_fact() function. The links to the source GitHub repositories are included in the comment line above each function.

The file "create\_tables.py" contains the functions responsible for reading the csv files created by the split\_csv() function and creating the 12 tables in the database.

# Description of Database Tables Columns and Handling of Missing/Unknown Values According to the Canadian Heart Health Survey Codebook

The CHH\_ProvincialHeartHealthData CSV file used in the project has 23129 rows (patient entries) with 266 columns for each row. The CSV file columns contained no missing/null values (screenshots below):

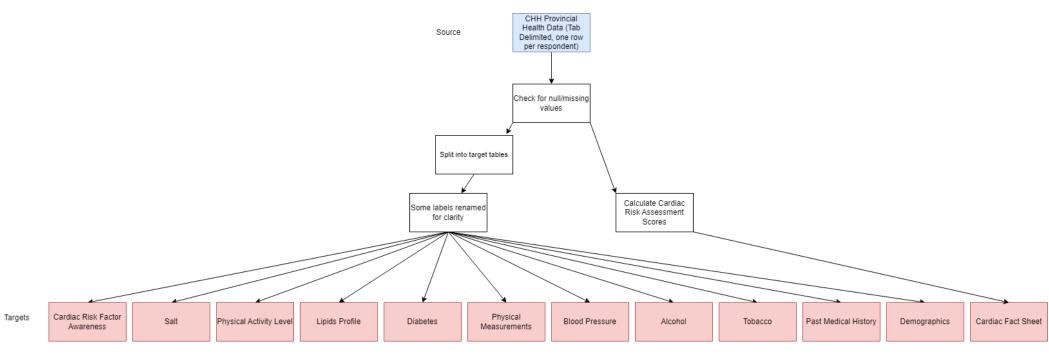
```
df = pd.read_csv('C:/users/conta/desktop/physical_design/CHH_ProvincialHeartHealthData.tab', encoding='ascii', sep='\t')
shape = df.shape
print("The shape of the data frame is "_shape)
print(df.isnull().values.any())
```

```
The shape of the data frame is (23129, 266)
False
```

All missing or unknown attributes in the CSV file were marked with special codes according to the survey's codebook. The tables below (after the Data Staging Plan) provide a description of columns in the database tables and the codes used for missing or unknown values. The SEQNO (Sequence Number) column in the CSV file (starting at 1, ending at 23129) was used as the primary key of the fact and dimensional tables. Screenshots of the tables created in the PostgreSQL database are also included in the report.

Due to this data being previously being made uniform and codified for a previous study, very little additional work was required to prepare it for use in our warehouse. Due to the structure of the survey questions useful data bins were already established within the dataset, some of which proved useful for our requirements.

## **Data Staging Plan**



#### **SALT Table**

Column Name	Description	Code description	Missing/Unknown Values
SEQNO	Sequence Number(Primary Key)	Primary Key	N/A
SALTCOOK	HOW OFTEN IS SALT ADDED WHILE COOKING	Often 1 Sometimes 2 Occasionally 3 Almost never or never 4	Variable unknown 7 Don't know 8 Not stated/refused 9
SALTFOOD	HOW OFTEN IS SALT ADDED AT THE TABLE	Often 1 Sometimes 2 Occasionally 3 Almost never or never 4	Variable unknown 7 Not stated/refused 9
SALTBP	SALT EFFECTS: BP WOULD INCREASE?	Mentioned 1 Not mentioned 2	Variable unknown 7 Not applicable 0
SALTHA	SALT EFFECTS: INCR RISK OF HRT ATTACK?	Mentioned 1 Not mentioned 2	Variable unknown 7 Not applicable 0
SALTSTR	SALT EFFECTS: INCREASE RISK OF STROKE?	Mentioned 1 Not mentioned 2	Variable unknown 7 Not applicable 0

#### **DEMOGRAPHICS Table**

Column Name	Description	Code description	Missing/Unknown Values
SEQNO	Sequence Number(Primary Key)	Primary Key	N/A
PROV	PROVINCE	Newfoundland 10 Prince Edward Island 11 Nova Scotia 12 New Brunswick 13 Quebec 24 Ontario 35 Manitoba 46 Saskatchewan 47 Alberta 48 British Columbia 59	N/A
SEX	SEX	Male 1 Female 2	N/A
AGE	AGE IN YEARS	AGE IN YEARS (int)	N/A
EDUC	GROUPED EDUCATION	Completed elementary school or less 1	Not stated/refused 9

INCOME	INCOME LEVEL	Some secondary schooling 2 Secondary school completed 3 University degree completed 4 Under \$12,000 1 \$12,000 - \$24,999 2 \$25,000 - \$49,999 3 \$50,000 and over 4	Not stated/refused 9
INCADEQ	INCOME ADEQUACY	Low income 1 Middle income 2 High income 3	Unknown income 9
EMPLOY	EMPLOYMENT STATUS	Full-time (35 hours or more a week) 1 Part-time (Less than 35 hours a week) 2 Unemployed 3 Laid off 4 Retired 5 Other 6 Homemaker 7 Student 8	Not stated/refused 99
WORKTYPE	WORK CLASSIFICATION	Professional 1 Clerical worker 2 Skilled foreman 3 Manager/official/proprietor 4 Sales worker 5 Non-skilled 6 Other work 7	Not stated/refused 99 Not applicable 0
MARITAL	CURRENT MARITAL STATUS	Never married 1 Divorced 2 Married/common law 3 Widowed/widower 4 Separated 5	Not stated/refused 9
HOUSEHLD	NUMBER OF PERSONS IN HOUSEHOLD	NUMBER OF PERSONS IN HOUSEHOLD	Not stated/refused 99
LANG	LANGUAGE	English 1 French 2 Other 3	Variable unknown 7 Not stated/refused 9

#### **ALCOHOL Table**

Column Name	Description	Code description	Missing/Unknown Values
SEQNO	Sequence Number(Primary Key)	Primary Key	N/A
ALCOHOL	ALCOHOL DRINKING STATUS	Never drink 1 Former drinker 2 Current drinker 3	Variable unknown 7 Not stated/refused 9
ALC12MTH	TAKEN DRINK IN PAST 12 MONTHS?	Yes 1 No 2	Variable unknown 7 Not stated/refused 9 Not applicable 0
ALCMTH	MONTHLY FREQ OF ALC DRINKS IN PAST YEAR	MONTHLY FREQ OF ALC DRINKS IN PAST YEAR (int) Less than once a month 95	Variable unknown 97 Don't know 98 Not stated/refused 99 Not applicable 0
ALCNUM	AVERAGE DAILY ALCOHOL CONSUMPTION	AVERAGE DAILY ALCOHOL CONSUMPTION (int)	Variable unknown 97 Don't know 98 Not stated/refused 99 Not applicable 0

#### **TOBACCO Table**

Column Name	Description	Code description	Missing/Unknown Values
SEQNO	Sequence	Primary Key	N/A
	Number(Primary Key)		
REGSMOK	REGULAR SMOKER?	No 0	Not stated/refused 9
		Yes 1	
SMOKECAT	CURRENT SMOKING	Never smoked 1	Not stated/refused 9
	STATUS	Former smoker 2	
		Regular cigarette	
		smoker 3	
		Occasional	
		cigarette smoker 4	
		Pipe or cigar	
		smoker 5	
CIGCAT	CIGARETTES SMOKED	Not a regular	Variable unknown 7
	CATEGORIES	cigarette smoker 1	Not stated/refused 9
		1-10 per day 2	
		11-25 per day 3	
		26 or more per day	
		4	

CIG_DAY	NUMBER CIGARETTES	NUMBER	Variable unknown 997
	PER DAY	CIGARETTES PER	Not stated/refused 999
		DAY (int)	Not stated/refused 99
PIPE	DO YOU SMOKE A PIPE	Yes 1	Not stated/refused 9
	NOW?	No 2	Not applicable 0
CIGAR	DO YOU SMOKE CIGARS	Yes 1	Not stated/refused 9
	NOW?	No 2	Not applicable 0

## PAST\_MEDICAL\_HISTORY Table

Column Name	Description	Code description	Missing/Unknown
			Values
SEQNO	Sequence Number(Primary	Primary Key	N/A
	Key)		
EVERHA	HAVE YOU EVER HAD A HEART	Yes 1	Not sure 8
	ATTACK?	No 1	Not stated/refused 9
EVERSTR	HAVE YOU EVER HAD A	Yes 1	Not sure 8
	STROKE?	No 2	Not stated/refused 9
OTHHD	DO YOU HAVE ANY OTHER	Yes 1	Not sure 8
	HEART DISEASE?	No 2	Not stated/refused 9
RXHEART	TAKING ANY PRESCRIBED	Yes 1	Not sure 8
	MEDS FOR HEART?	No 2	Not stated/refused 9
FEMRX	ON ORAL	Oral	Not stated/refused 9
	CONTRACEPTIVES/HORMONAL	contraceptives 1	Not applicable 0
	PILLS?	Hormonal pills 2	
		Neither 3	

## **BLOOD\_PRESSURE Table**

Column Name	Description	Code description	Missing/Unknown Values
SEQNO	Sequence Number(Primary Key)	Primary Key	N/A
SYSCAT	SYSTOLIC BLOOD PRESSURE CATEGORIES	Less than 120 1 120-129 2 130-139 3 140-149 4 150-159 5 160 or more 6	Not stated/refused 9
DIASCAT	DIASTOLIC BLOOD PRESSURE CATEGORIES	Less than 80 1 80-84 2 85-89 3	Not stated/refused 9

	T	100.04.4	
		90-94 4	
13/050	LIVE ETTENSIVE (LIDE) STATUS	100 or more 6	1 21/2
HYPER	HYPERTENSIVE (HBP) STATUS	Not hypertensive	N/A
		0	
		Hypertensive 1	<u> </u>
HATCS	HBP AWARE TREATMENT	Treated and	Variable unknown 7
	CONTROL STATUS	controlled 1	Not applicable (HYPER =
		Treated and not	0) 0
		controlled 2	
		Not treated and	
		not controlled 3	
		Unaware 4	
PHARM	ON PHARMACOLOGICAL	No 0	N/A
	TREATMENT FOR HBP?	Yes 1	
BPSALT2	BP TREATMENT NOW: SALT	Mentioned 1	Not applicable 0
	FREE DIET?	Not mentioned 2	
BPWGT2	BP TREATMENT NOW: WATCH	Mentioned 1	Not applicable 0
	WEIGHT?	Not mentioned 2	
BPSTR2	BP TREATMENT NOW: AVOID	Mentioned 1	Not applicable 0
	STRESS?	Not mentioned 2	
BPSMOK2	BP TREATMENT NOW:	Mentioned 1	Not applicable 0
	REDUCE/STOP SMOKING?	Not mentioned 2	
BPALC2	BP TREATMENT NOW: REDUCE	Mentioned 1	Not applicable 0
	ALCOHOL?	Not mentioned 2	
BPEXER2	BP TREATMENT NOW: START	Mentioned 1	Not applicable 0
	EXERCISE?	Not mentioned 2	
BPSALT	HBP RELATED TO: SALT?	Mentioned 1	Not stated/refused 9
		Not mentioned 2	Not applicable 0
BPFATS	HBP RELATED TO: FATS?	Mentioned 1	Not stated/refused 9
		Not mentioned 2	Not applicable 0
BPCALOR	HBP RELATED TO:	Mentioned 1	Not stated/refused 9
	CALORIES/OVEREATING?	Not mentioned 2	Not applicable 0
BPALC	HBP RELATED TO: ALCOHOL?	Mentioned 1	Not stated/refused 9
-		Not mentioned 2	Not applicable 0
BPCAFF	HBP RELATED TO:	Mentioned 1	Not stated/refused 9
	CAFFEINE/COFFEE?	Not mentioned 2	Not applicable 0
BPHEALTH	DO YOU THINK HIGH BP CAN	Yes 1	N/A
	AFFECT HEALTH?	No 2	.,
		Do not know 8	
		1 20 HOURIOW 0	1

## PHYSICAL\_MEASUREMENTS Table

Column Name	Description	Code description	Missing/Unknown Values
SEQNO	Sequence Number(Primary Key)	Primary Key	N/A
HGTC	CLINIC: HEIGHT IN CENTIMETRES	HEIGHT IN CENTIMETRES (int)	Variable unknown 997 Recoded missing 998 Not stated/refused 999
WGTC	CLINIC: WEIGHT IN KILOGRAMS	WEIGHT IN KILOGRAMS (int)	Variable unknown 997 Recoded missing 998 Not stated/refused 999
BMI	BODY MASS INDEX	BODY MASS INDEX (int)	Variable unknown 999 Not stated/refused 999
BMICAT	BMI CATEGORIES	Less than 20 1 20-24.9 2 25-26.9 3 27-29.9 4 30-34.9 5 35 or more 6	Variable unknown 7 Not stated/refused 9
WAIST	CLINIC: WAIST IN CENTIMETRES	WAIST IN CENTIMETRES (int)	Not stated/refused 997 Variable Unknown 999
HIP	CLINIC: HIP IN CENTIMETRES	HIP IN CENTIMETRES (int)	Not stated/refused 997 Variable Unknown 999
WHR	WAIST/HIP RATIO	WAIST/HIP RATIO (int)	Not stated/refused 99 Variable Unknown 99
WHRCAT	WAIST/HIP RATIO RISK CATEGORIES	Low waist/hip ratio 0 High waist/hip ratio 1	Variable unknown 7 Not stated/refused 9
PULSE	CLINIC: PULSE MEASUREMENT	PULSE MEASUREMENT (int)	Variable unknown 997 Not stated/refused 999

#### **DIABETES Table**

Column Name	Description	Code description	Missing/Unknown
			Values
SEQNO	Sequence Number(Primary Key)	Primary Key	N/A

DIABET	EVER TOLD YOU HAVE	Yes 1	Variable unknown 7
	DIABETES?	No 2	Not stated/refused 9
DIABAGE	AGE WHEN TOLD YOU HAD	AGE WHEN TOLD	Variable unknown 97
	DIABETES	YOU HAD	Can't remember 98
		DIABETES (int)	Not stated/refused 99
			Not applicable 0
DBNOTRT	DIABETES TX NOW: NO	Mentioned 1	Variable unknown 7
	CURRENT TREATMENT?	Not mentioned 2	Not stated/refused 9
			Not applicable 0
DBINS	DIABETES TX NOW: INSULIN?	Mentioned 1	Variable unknown 7
		Not mentioned 2	Not stated/refused 9
			Not applicable 0
DBRX	DIABETES TX NOW: PILLS TO	Mentioned 1	Variable unknown 7
	CONTROL B SUG?	Not mentioned 2	Not stated/refused 9
			Not applicable 0
DBDIET	DIABETES TX NOW: DIET?	Mentioned 1	Variable unknown 7
		Not mentioned 2	Not stated/refused 9
			Not applicable 0
DBWG	DIABETES TX NOW: WEIGHT	Mentioned 1	Variable unknown 7
	LOSS?	Not mentioned 2	Not stated/refused 9
			Not applicable 0
DBDK	DIABETES TX NOW: DONT	Mentioned 1	Variable unknown 7
	KNOW/NOT STATED	Not mentioned 2	Not stated/refused 9
			Not applicable 0

## LIPIDS\_PROFILE Table

Column Name	Description	Code description	Missing/Unknown
			Values
SEQNO	Sequence Number(Primary	Primary Key	N/A
	Key)		
TCHOL	TOTAL PLASMA CHOLESTEROL	TOTAL PLASMA	Variable unknown
	(MMOL/L)	CHOLESTEROL	999.97
		(float)	No measurement 999.99
TCHOLCAT	TOTAL CHOL CATEGORIES	Less than 5.2 1	Did not attend clinic or
		5.2 - 6.1 2	no blood sample 7
		6.2 or more 3	No measurement 9
HDL	HIGH DENSITY LIPOPROTEINS	PLASMA HIGH	Variable unknown
	(MMOL/L)	DENSITY	999.97
		LIPOPROTEINS	No measurement 999.99
		(MMOL/L) (float)	
LDL	LOW DENSITY LIPOPROTEINS	PLASMA LOW	Variable unknown
	(MMOL/L)	DENSITY	999.97
			No measurement 999.99

		LIPOPROTEINS (MMOL/L) (float)	
TRIG	TRIGLYCERIDES (MMOL/L)	PLASMA TRIGLYCERIDES (MMOL/L) (float)	Variable unknown 999.97 No measurement 999.99
BCHRX	TO LOWER CHOL: TAKE PRESCRIBED MEDS?	Mentioned 1 Not mentioned 2	Not applicable 0
BCHFAT	TO LOWER CHOL: EAT LESS FATTY FOODS?	Mentioned 1 Not mentioned 2	Not applicable 0
BCHWGT	TO LOWER CHOL: LOSE WEIGHT?	Mentioned 1 Not mentioned 2	Not applicable 0
CHOLBP	CHOL AFFECTS HEALTH: INCREASE BP?	Mentioned 1 Not mentioned 2	Not stated/refused 9 Not applicable 0
CHOLHA	CHOL AFFECTS HEALTH: HEART ATTACK?	Mentioned 1 Not mentioned 2	Not stated/refused 9 Not applicable 0

# PHYSICAL\_ACTIVITY\_LEVEL Table

Column Name	Description	Code description	Missing/Unknown Values
SEQNO	Sequence Number(Primary Key)	Primary Key	N/A
SEDENT	SEDENTARY?	Yes 1 No 0	Not stated/refused 9
EXER	EXERCISE REGULARLY (1+ TIMES A WEEK)?	Yes 1 No 2	Cannot remember 8 Not stated/refused 9
EXERSTRN	HOW MUCH OF THIS EXERCISE IS STRENUOUS?	Most of it 1 Some of it 2 None of it 3	Not stated/refused 9 Not applicable 0
EXERLONG	HOW LONG DO YOU USUALLY EXERCISE?	Less than 15 minutes 1 15 to 30 minutes 2 31 to 60 minutes 3 More than 60 minutes 4 Don't know 8	Not stated/refused 9 Not applicable 0
EXERWORK	WORK REQUIRES STRENUOUS PHYS ACTIVITY?	Yes 1 No 2	Variable unknown 7 Cannot remember 8 Not stated/refused 9 Not applicable 0

## CARDIAC\_RISK\_FACTORS\_AWARENESS Table

Column Name	Description	Code description	Missing/Unknown Values
SEQNO	Sequence Number(Primary Key)	Primary Key	N/A
HDPREV	BELIEVE HEART DISEASE CAN	Yes 1	Not sure 8
	BE PREVENTED?	No 0	Not stated/refused 9
HDDIET	CAUSES OF HRT DISEASE:	Mentioned 1	N/A
	POOR DIET?	Not mentioned 2	
HDWGT	CAUSES OF HRT DISEASE:	Mentioned 1	N/A
	OVERWEIGHT?	Not mentioned 2	
HDFATS	CAUSES OF HRT DISEASE:	Mentioned 1	N/A
	EXCESS FATS?	Not mentioned 2	
HDSALT	CAUSES OF HRT DISEASE:	Mentioned 1	N/A
	EXCESS SALT?	Not mentioned 2	
HDCHOLBD	CAUSES OF HRT DISEASE: HIGH	Mentioned 1	N/A
	BLOOD CHOL?	Not mentioned 2	
HDCHOLFD	CAUSES OF HRT DIS: FOODS	Mentioned 1	N/A
	WITH HIGH CHOL?	Not mentioned 2	
HDSTRESS	CAUSES OF HRT DISEASE:	Mentioned 1	N/A
	STRESS/WORRY?	Not mentioned 2	
HDTIRED	CAUSES OF HRT DISEASE:	Mentioned 1	N/A
	OVERWORK/FATIGUE?	Not mentioned 2	
HDEXER	CAUSES OF HRT DISEASE: LACK	Mentioned 1	N/A
	OF EXERCISE?	Not mentioned 2	
HDSMOKE	CAUSES OF HRT DISEASE:	Mentioned 1	N/A
	SMOKING?	Not mentioned 2	
HDHERED	CAUSES OF HRT DISEASE:	Mentioned 1	N/A
	HEREDITY?	Not mentioned 2	
HDHBP	CAUSES OF HRT DISEASE: HIGH	Mentioned 1	N/A
	BP?	Not mentioned 2	
HDART	CAUSES OF HRT DISEASE:	Mentioned 1	N/A
	ARTERIOSCLEROSIS?	Not mentioned 2	
HDDK	CAUSES OF HRT DISEASE: DO	Mentioned 1	N/A
	NOT KNOW?	Not mentioned 2	

## CARDIAC\_FACT Table

Column Name	Description	Missing/Unknown Values
SEQNO	Sequence Number(Primary Key)	N/A

DEMOGRAPHICS_KEY	DEMOGRAPHICS Table FK	N/A
PHYSICAL_MEASUREMENTS_KEY	PHYSICAL_MEASUREMENTS Table FK	N/A
BLOOD_PRESSURE_KEY	BLOOD_PRESSURE Table FK	N/A
LIPIDS_PROFILE_KEY	LIPIDS_PROFILE Table FK	N/A
DIABETES_KEY	DIABETES Table FK	N/A
ALCOHOL_KEY	ALCOHOL Table FK	N/A
TOBACCO_KEY	TOBACCO Table FK	N/A
SALT_KEY	SALT Table FK	N/A
PHYSICAL_ACTIVITY_LEVEL_KEY	PHYSICAL_ACTIVITY_LEVEL Table FK	N/A
PAST_MEDICAL_HISTORY_KEY	PAST_MEDICAL_HISTORY Table FK	N/A
CARDIAC_RISK_FACTORS_AWARENESS_K	CARDIAC_RISK_FACTORS_AWARENE	N/A
EY	SS Table FK	
FRAMINGHAM_SCORE (Calculated from	The Framingham Coronary Heart	Missing
Dimension Tables' Attributes)	Disease Risk Score estimates risk of	Parameters/Unabl
	heart attack in 10 years.	e to Calculate 999
ASCVD_SCORE (Calculated from	The ASCVD (atherosclerotic	Missing
Dimension Tables' Attributes)	cardiovascular disease) risk score	Parameters/Unabl
	estimates risk of heart attack in 10	e to Calculate 999
	years	

#### **Database Tables Screenshots**

