|  |  |
| --- | --- |
| DATA STRUCTURES -Mental health in tech survey | Abstract  This dataset is from a 2014 survey that measures attitudes towards mental health and frequency of mental health disorders in the tech workplace  Monika Pawar  [MS-AIM] |

*Project Phase 1 & 2*

1. Briefly describe what the data set you identified is about, what file format it came in, and what the source of the data set is. If you downloaded the data from an online repository, note that this is a repository of data from other sources, so be sure to read the documentation regarding what the original source of the data was and note both the repository and the original source.

Dataset: **Mental Health in Tech Survey (2014)**

This dataset is about the Survey on Mental Health in the Tech Workplace in 2014, which measures attitudes towards mental health and frequency of mental health disorders in the tech workplace. I used this dataset from <https://www.kaggle.com/osmi/mental-health-in-tech-survey> and it come in the CSV format.

The original dataset is from Open Sourcing Mental Illness (OSMI) and can be downloaded here. (<https://osmihelp.org/research>)

Description: With over 1200 responses, it was believed that the 2014 Mental Health in Tech Survey was the largest survey done on mental health in the tech industry. Since then, OSMI has conducted two more surveys, 2016 and 2017.

2. Prepare a set of data names and definitions for all terms related to the data. For example, if your data set focuses on movies and what actors appear in each movie, define what a movie is, what an actor is, and all other significant terms related to the data.

**Survey**: An examination of opinions, behavior, etc., made by asking people number of questions.

**Participant**: Person who takes part in the survey

**Place**: Any location in a country where the participant lives.

3. Prepare a metadata table (in Word) of attributes available in the data set. See Table 1-1 on pg. 8 of Hoffer et al textbook for a sample.



4. Draw a logical model of the data set as it currently exists, including all tables and fields. Identify relationships between the tables if they current exist.



5. After considering the data, draw a normalized logical model (through 3NF) that could be used to store the data in your data set. Create lookup tables for all categorical variables as well. Document assumptions as needed.

* Full Functional Dependency

Participant\_id, Survey\_Id 🡪 Timestamp, Age, Gender, Country ,state, self\_employed, family\_history, treatment, work\_interfere, no\_employees, remote\_work, tech\_company, benefits, care\_options, wellness\_program, seek\_help, anonymity, leave, mental\_health\_consequence, phys\_health\_consequence, coworkers, supervisor, mental\_health\_interview, phys\_health\_interview, mental\_vs\_physical, obs\_consequence, comments

* Partial dependencies

Participant\_id 🡪 Age, Gender, Country, state, timestamp

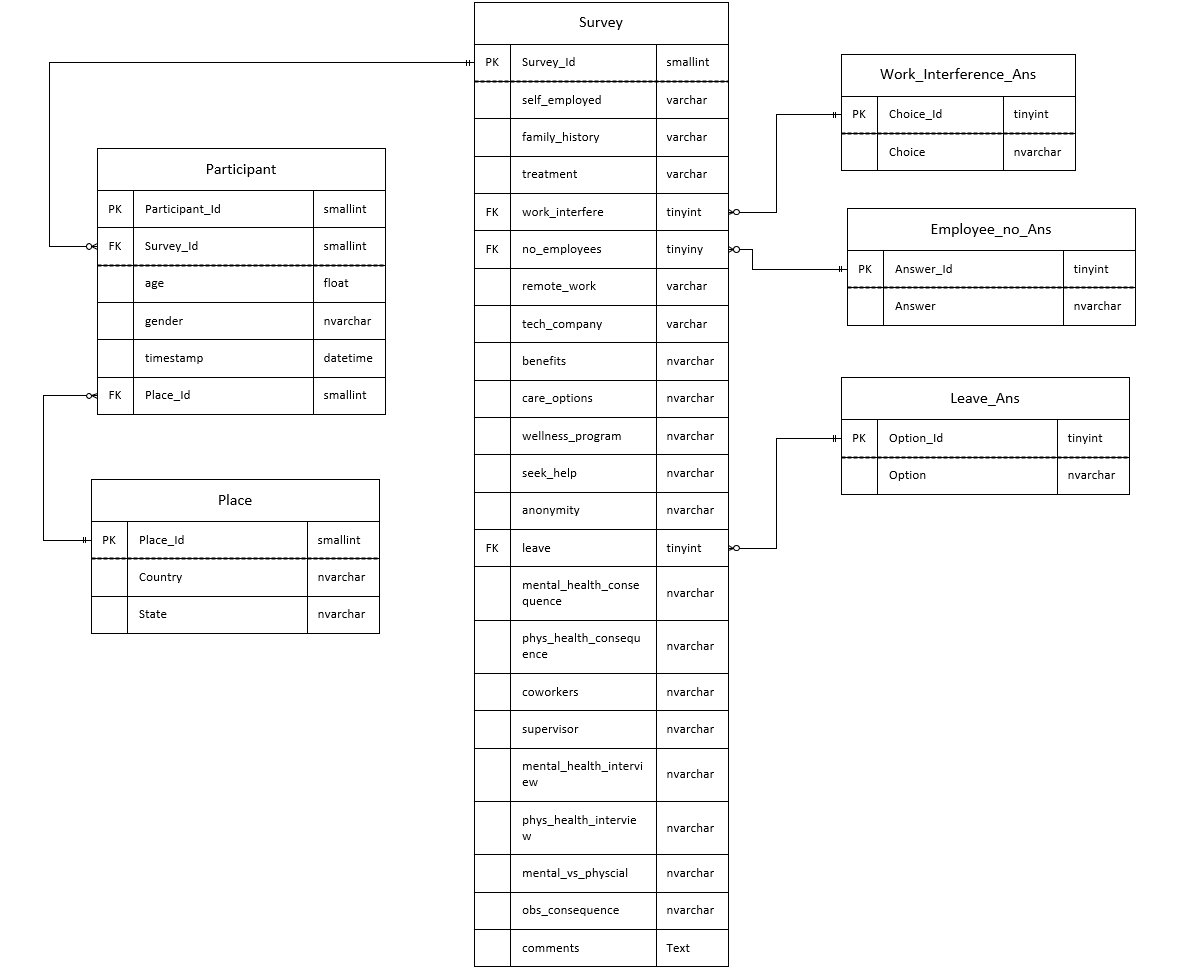
* Transitive dependencies

Place\_id 🡪 Country, State

* Lookup tables

Work\_Interference\_Ans, Employee\_Ans, Leave\_Ans

Logical Model with Normalization



6. Describe at least three questions that you can potentially answer using the data set.

* Are mental health illnesses more frequent among Tech workers as compared to non-Tech?
* How does the size of company relate to an employer formally discussing mental health?
* How does Age relate to comfort discussing mental health issues with peers?
* How does the frequency of mental health illness and attitudes towards mental health vary by geographic location?

*Project Phase 3*

A. Design of a dimensional model (star schema) for your chosen data set.

Using 2 of the questions you identified in phase 1 as the starting point and driver for your data mart, draw a dimensional model (star schema) that could be implemented to support answering these questions. Show your work by documenting the following:

1. Draw a fact-qualifier matrix (see Hoffer et al page 420 for an example).

**NOTE**: ‘treatment’ is a column which ask question that 🡪 ‘Have you sought treatment for a mental health condition?’ Here 1 means yes, participant sought treatment for a mental health condition and 0 means No, they have not sought treatment for a mental health condition.

|  |  |  |  |
| --- | --- | --- | --- |
| 1. How does the size of company relate to an employer formally discussing mental health? 2. How does Age relate to comfort discussing mental health issues with peers? 3. How does the frequency of mental health illness and attitudes towards mental health vary by geographic location?  4.Are mental health illnesses more frequent among Tech workers as compared to non-Tech? | Participant | Company | Count of Treatment (=1) |
|  |  |  |  |
| Participant Age | 2 |  |  |
| Company Size |  | 1 |  |
| Company Type |  |  | 4 |
| Country |  |  | 3 |
| State |  |  | 3 |

2. Provide a list of the necessary dimensions and a brief description of each. Note whether any of the dimensions are hierarchies or multivalued.

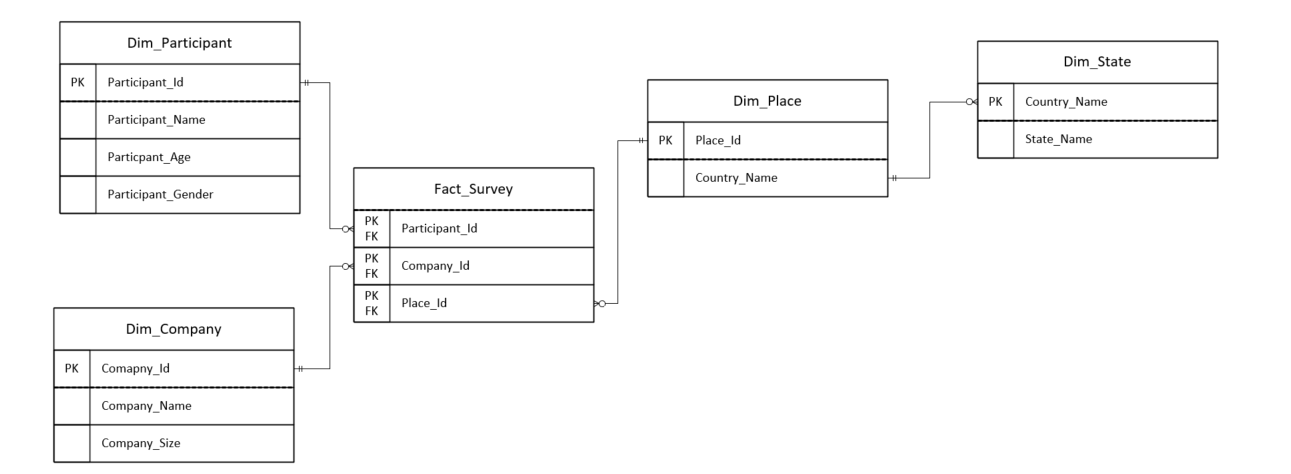
1. Participant Id: A unique Id number for participant who is taking the Mental health survey.
2. Participant Age: Age of the participant who takes the mental health survey.
3. Company Size: The size of the company in which the participant works.
4. Country: Name of the country from where participants belongs.
5. State: Name of the state from where participants belongs.

In the data set, **Place** dimension is having the **Hierarchies** 🡪 country 🡪 state

3. Specify whether the fact table will be fact or factless and identify the grain of the fact table.

As per my understanding the fact table will be factless and the grain of the fact table will be Transactional Grain.

4. Draw a dimensional model (star schema) showing all tables and fields.



B. Briefly discussion the three data structures drawn for your project data set – the original, the logical normalized model, and the star schema. Compare and contrast the structures.

* Original dataset has a simple one table model in which it has all the details about the participant, all the questions asked in the survey and their answers given by the participants. This model is not normalized having all the details in one table.
* In the Logical model the dataset is normalized by following the rule till 3rd normal form, thus removed partial and transitive dependency. The look up tables has been made for the Answer key of various question asked in the survey. This logical model contains data in more organized way.
* The star schema model- this model contains the fact and dimensional tables for the questions we intended to get answer from the dataset. We have dimensional hierarchies for place which include country and state

C. Analyze the data to answer questions identified in phase 1.

* Using the data ingested in phase 2, select 2 questions you identified in phase 1 and prepare queries that allow you to answer the questions

***Ques 1) Are mental health illnesses more frequent among Tech workers as compared to non-Tech?***

NOTE:

Survey Question 🡪 **treatment**: Have you sought treatment for a mental health condition?

* + **1**: yes the participant has sought treatment for mental health condition

Survey Question🡪 **tech\_company**: Is your employer primarily a tech company/organization?

USE [MentalHealth\_Survey]

Select ST. tech\_company,count([treatment]) As 'No\_MH\_illness'

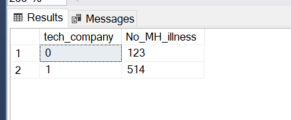
from Participant\_Table as PT

Join Survey\_Table as ST

On PT.Survey\_Id = ST.Survey\_Id

where ST.treatment = 1

Group by ST.tech\_company



We can see from the results that Mental health illness is more mental health illnesses more frequent among Tech workers as compared to non-Tech

***Ques 2) How does the frequency of mental health illness and attitudes towards mental health vary by geographic location?***

Survey Question 🡪 **treatment**: Have you sought treatment for a mental health condition?

Considering country and state of a participant as for geographical location

Select PlT.Country,PLT.state,count([treatment]) As 'No\_MH\_illness'

from Participant\_Table as PT

Join Survey\_Table as ST

On PT.Survey\_Id = ST.Survey\_Id

join Place\_Table as PlT

on PT.Place\_Id = PlT.Place\_Id

where ST.treatment = 1

Group by PlT.Country, PLT.state

Order by PlT.Country,PLT.state Desc

