Cadigoy. Dean Gabriel G.

CIT401A - IT31S1

1.) What challenges did you experience during the MySQL data-dump restoration?

　　During the process of dumping my data, I encountered an issue regarding the size as I came to realize the astounding size of 45k+ objects/rows, this was so far the biggest one I handled in my years of studying. It took me some time to configure certain items in xammp such as its ini file as well as other stuff regarding the sql in order to handle such a massive file.

2.) With respect to the newly created MySQL table, complete the details below:

Number of Attributes:	18
Number of Objects:	45,211

3.) With respect to the objects inside the table, complete the details below:

Attribute	Data Type
ID	Integer
Age	Integer
Job	Varchar
Marital	Varchar
Education	Varchar
Default_Status	Varchar
Balance	Decimal
Housing	Varchar
Loan	Varchar
Contact	Varchar
Day	Tiny Integer
Month	Varchar
Duration	Integer
Campaign	Integer
Pdays	Integer
Previous	Integer
Outcome	Varchar
Target	Varchar

Job Attribute

Initial Summary		Final Summary	
job	**job_count** ▾ 1	**job**	**people_count** ▾ 1
blue-collar	9732	blue-collar	9732
management	9458	management	9458
technician	7597	technician	7597
admin.	4996	admin.	5171
services	3639	services	4154
retired	2170	retired	2264
self-employed	1579	self-employed	1579
entrepreneur	1487	entrepreneur	1487
unemployed	1303	unemployed	1303
housemaid	1240	housemaid	1240
student	938	student	938
svcs	515	unknown	288
unknown	277		
Adm	175		
Ret	94		
???	11		

Marital Attribute

Initial Summary		Final Summary	
marital	**marital_count** ▾ 1	**marital**	**people_count** ▾ 1
married	27214	married	27214
single	12790	single	12790
divorced	4624	divorced	5207
Div	583		

Education Attribute

Initial Summary		Final Summary	
education	**education_count** ▾ 1	**education**	**people_count** ▾ 1
secondary	23082	secondary	23202
tertiary	13272	tertiary	13301
primary	6851	primary	6851
unknown	1843	unknown	1857
2nd	120		
3rd	29		
???	14		

Contact Attribute

Initial Summary		Final Summary	
contact	**contact_count** ▾ 1	**contact**	**people_count** ▾ 1
cellular	28947	cellular	29285
unknown	13020	unknown	13020
telephone	2719	telephone	2906
cell.	338		
Tel	187		

Based on the results produced in items #1 and #2, what issues did you encounter during the clean-up process?

       During the cleanup process, finding the right SQL query can become an issue especially when different datatypes are used. The data cleanup is overall simple, but can become complex upon certain conditions during the process of constructing the SQL query.

What are the possible repercussions if data clean-up is neglected during the "Data Preparation" phase of the CRISP-DM framework?

       During my observation on cleaning up the data, especially around the contact part, I observed that the entire first few rows are unknowns and if I did not clean the data beforehand, I might come to see that unknown may be the most common instead of contact which is the actual most common. I have also observed that clean data is much easier to read and much more manageable leading to having it easier to understand causing less errors.