424 Homework 1

Marc Wells

January 6, 2016

1.

We too often use spreadsheets to do more than the most simple statistical procedures, which often results in erroneous results and outcomes. This is dangerous because Excel and other spreadsheets were not made to run statistics or be a database, but that is often how people use them.

2.

This set of readings is essentially about the basics of database design and manipulation. It dealt with primary and foreign keys, simple datasets to many-to-many datasets, rules for normalization, and database management languages.

3.

I have read this

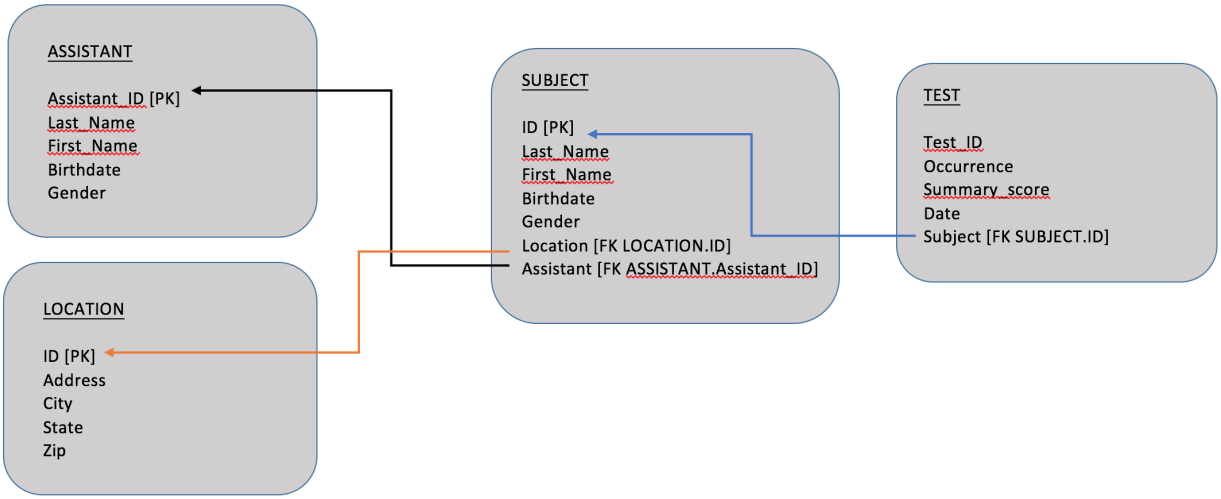
4.

```
city <- c("Bath","Blackburn","Derby","Southhampton","York","Oxford")
mortality <- c(1247,1800,1555,1369,1378,1175)
calcium <- c(105,14,39,68,71,107)
air <- data.frame(city,mortality,calcium)
air
```

```
##           city mortality calcium
## 1         Bath      1247     105
## 2    Blackburn      1800      14
## 3        Derby      1555      39
## 4 Southhampton      1369      68
## 5         York      1378      71
## 6       Oxford      1175     107
```

5.

The database scheme would follow the pattern below.



Test_ID is the primary key in the test table.

In order to keep the first normal form, we ensured that there were no duplicate columns and that information could not hold more than one value. To keep the second normal form, we split the tables up in such a way that the primary key related to every other variable in the table. The "test" table has subject_id as the foreign key, and some people might argue that person has nothing to do with the test, but in this case, the subject taking the test is an important part of the test scores and other metrics. Finally we kept the third normal form by having all variables describe only the primary key of each table. Subject in the "test" table describes test_ID by saying who took test number 4, for example.