Databases Project Phase 1: Proposal

Dongrui Zhong, Bowen Li

November 19, 2015

1. Team

Dongrui Zhong(Dzhong3), Bowen Li(bli26).

2. Target Domain

Economical data ranges the whole United States. Data includes people size, different kinds of ratios(including education ratio, sex ratio age ratio etc.), median income, employments of all states and different years.

3. Questions

1. List all states with the Health ratios with ChildrenPerHousehold greater and equal than two;

2. List the GDP and Agriculture, Industry ratio and AnnualPayroll of the states with top 10 highest numbers of HighSchools.

3. List the PerCapitaIncome, MedianHouseholdIncome of all state with all information in the Families table, with the constraint that Male Sex ratio is greater than the Female ratio.

4. List the Income information like in 3 of all state with the constraint that Male Sex ratio is smaller and equal than the Female ratio.

5. List the DiplomaRate of Bachelor, Master, Doctor of states which has top 10 highest numbers of HighSchool. Also, together list all the Income information and Business information of these states.

6. List the Race, Income and Health information of all states after(not equal than) Year 2008.

7. List some statistics of Employment information of all states of every year. The statistics include sample mean, sample deviation, sample median.

8. List the mean of sum of Hospitals of all states, grouped by ChildrenPerHousehold(rounded values). Together show the median MedianHouseholdIncome of them.

9. List the population ratios of states together with the Employment ratios of them.

10. List the HomeOwnershipRate of states with Families who have ChildrenPerHousehold greater and equal than 2, together with all information in the Income table.

11. List the information in Health table together with Business information of states, which has a greater Female rate than Male after Year 2005.

12. List all Industry and Services rates together with Race ratios of all states which has greater number of Male than Female with top 10 greatest ratio MedianHouseholdIncome.

13. List all information of Economy, Business, Health tables together with PerCapitaIncome of all states. These states should have ChildrenPerHousehold greater and equal than 1, and the sum ratio of DiplomaRate of higher than Bachelor is greater than 32%;

14. List the Poverty rate of all states which is of top 10. Also List number of HighSchool the their number ranks(ordered by increasing number) of these states.

15. List the InsuranceCoverage of all states together their Race ratios of top 10 number of NumberOfFirms.

4. Relational Data Model

Population	State	Year	0-5	5-17	18-64	65+	Total
	MD	2014	6.20%	16.40%	63.60%	13.80%	5976407

Sex	State	Year	Male	Female
	MD	2014	48.50%	51.50%

Race	State	Year	White	AfericanAmerican	Asian	Hispanic	NativeAmerican	Other
	MD	2014	52.60%	30.30%	6.40%	9.30%	0.60%	0.80%

Economy	State	Year	GDP(Millions)	Agriculture	Industry	Services
	MD	2014	$348,631	0.28%	14.07%	85.65%

Business	State	Year	FirmSize	NumberOfFirms	NumberOfEmployees	AnnualPayroll($1000)
	MD	2011	20-99	10314	385371	17132816

EmploymentByAgeGroup	State	Year	14-18	19-21	22-24	25-34	35-44	45-54	55-64	65+	UnemploymentRate
	MD	2014	47821	98695	147861	523376	499579	559512	415764	152423	5.80%

EducationLevel	State	Year	HighSchoolOrHigher	BachelorOrHigher
(Age25+)	MD	2009	88.20%	35.70%

NumSchools	State	Year	University&College	HighSchool
	MD	2014	55	248

Health	State	Year	InsuranceCoverageRate	NumberOfHospitals	MedicalCareRevenue
	MD	2014	94%	80	$52,337,215

Housing	State	Year	NumHousingUnits	BuildingPermits	HomeOwnershipRate
	MD	2014	2422194	16331	67.60%

Families	State	Year	PersonPerHousehold	ChildrenPerHousehold
	MD	2014	2.65	0.88

Income	State	Year	PerCapitaIncome	MedianHouseholdIncome	PovertyRate
	MD	2014	$36,354	$73,538	9.80%

5. SQL Statement

Here we list several statements from problem 3.

```
1.
SELECT h.*
FROM HEALTH AS h, Families AS f
WHERE f.ChildrenPerHousehold >= 2 AND f.State = h.State AND f.Year = h.Year;

2.
SELECT e.*
FROM Economy AS e,
     (SELECT od.*
      FROM (SELECT d.*
            FROM DiplomaRate AS d
            ORDER BY d.HighSchool) AS od
      LIMIT 10) AS temp
WHERE  AND e.State = temp.State AND e.Year = temp.Year;


3.
SELECT i.PerCapitaIncome, i.MedianHouseholdIncome, f.*
FROM Income AS i, Families AS f, Sex AS s
WHERE s.Male > s.Female AND i.State = f.State AND i.Year = f.Year AND f.State =
    s.State AND f.Year = s.Year;

4.
SELECT i.PerCapitaIncome, i.MedianHouseholdIncome, f.*
FROM Income AS i, Families AS f, Sex AS s
WHERE s.Male < s.Female AND i.State = f.State AND i.Year = f.Year AND f.State =
    s.State AND f.Year = s.Year;
```

6. How to Load Values

We will load data from the following website: https://census.gov and get the ".csv" or some other data type and load them into our own database.

7. Report

The final report can be a small website including frontend, backend and our database. Using the Model-View-Controller design pattern to design an interface satisfying people viewing some statistics of data, access of the backend whole or part dataset when login via password. Also, for a great part of the interface, we may implement some data mining algorithms to do some prediction using data in the database.

8. Advanced Topics

We may focus mainly on designing data mining(linear regression model) algorithms and implementing an full stack website(interface) using Python and Django.

9. Database Platform

We may implement using Python and Django, and using SQLite as our Database. One of our computer is Macbook(1.1 GHz Intel Core M, 8 GB 1600 MHz DDR3, Intel HD Graphics 5300 1536 MB) and anther is Lenovo(Intel(R) Core(TM) i7-4510U CPU 2.00 GHz 2.60 GHz, 16GB RAM)