STAT 4260 – Databases
Spring 2018

Assignment 7 – [27 points]

[2pts] Electronic submission

1. **[8pts] Find the average accident severity and count for different types of motorcycles using a SQL statement**. Use column aliases, table aliases.

This uses the Vehicles_2015.csv, Accidents_2015.csv, and Road-Accident-Safety-Data-Guide.xlsx data. Note these are large files and you will need to use Python to import the data correction into a SQL server if you want to test your code. Therefore, you can assume the following. The data is stored in three tables (subsections of the outputs are shown below). Accidents 2015 and vehicles 2015 share the accident index field. And vehicles 2015 and vehicle type share the vehicle code field.

```
CREATE TABLE accidents_2015 (
Accident_index VARCHAR(13)
Accident_severity INT);
```

accident_index	accident_severity
201501BS70001	3
201501BS70002	3
201501BS70004	3
201501BS70005	3
201501BS70008	2
201501BS70009	3
201501BS70010	3
201501BS70011	3
201501BS70012	3

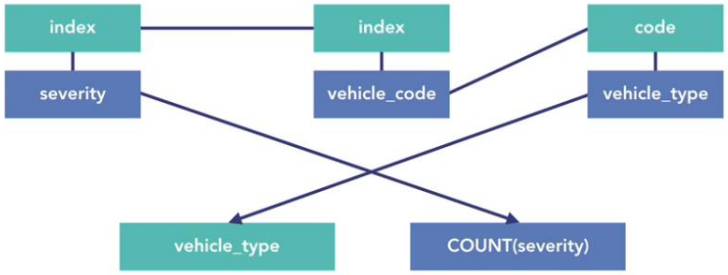
```
CREATE TABLE vehicles_2015 (
Accident_index VARCHAR(13)
vehicle_type VARCHAR(10));
```

accident_index	vehicle_type
201501BS70001	19
201501BS70002	9
201501BS70004	9
201501BS70005	9
201501BS70008	1
201501BS70008	9
201501BS70009	3
201501BS70009	19
201501BS70010	9

```
CREATE TABLE vehicle_type (
vcode INT,
vtype VARCHAR(100));
```

vcode	vtype
1	Pedal cycle
2	Motorcycle 50cc and under
3	Motorcycle 125cc and under
4	Motorcycle over 125cc and up to 500cc
5	Motorcycle over 500cc
8	Taxi/Private hire car
9	Car
10	Minibus (8 - 16 passenger seats)
11	Bus or coach (17 or more pass seats)

Diagram with connections (use the names in the CREATE TABLEs not from this diagram.



The output you are trying to should be similar to the following setup (without the grand total and you can use different column names):

Row Labels	Average of Accident_Severity	Count of Accident_Severity
Electric motorcycle	2.444444444	9
Motorcycle - unknown cc	2.694545455	275
Motorcycle 125cc and under	2.780701754	9234
Motorcycle 50cc and under	2.82655342	2237
Motorcycle over 125cc and up to 500cc	2.69044353	2187
Motorcycle over 500cc	2.58491636	7054
Grand Total	**2.709135073**	**20996**

2. **[2pts] Describe, in complete sentences, what the provided output for Question 1 informs us.**

3. **[5pts] Create a list of male and female populations for each county in California in 2014 and to return that information in the following formatted table using a SQL statement** (no need to worry about the 1000 separator (,) in the numbers). Hint, you may want to use a subquery.

Year: 2014	Male	Female
Alameda	785,307	814,288
Alpine	599	564
Amador	19,893	17,544
…	…	…

This uses data from CA_DRU_project_2010-2016.csv, a very large data set. Again, you will need to use Python to import the data correction into a SQL server if you want to test your code. Therefore, you can assume the following. Subsection of table output shown

```
CREATE TABLE pop_proj (
  county_code INT,
  county` VARCHAR(45),
  date_year VARCHAR(4),
  race_code INT,
  race TEXT,
  age INT,
  population INT);
```

county_code	county	date_year	race_code	race	gender	age	population
6001	Alameda	2010	1	White. Non-Hispanic	Female	0	2078
6001	Alameda	2010	1	White. Non-Hispanic	Female	1	2038
6001	Alameda	2010	1	White. Non-Hispanic	Female	2	2008
6001	Alameda	2010	1	White. Non-Hispanic	Female	3	2129
6001	Alameda	2010	1	White. Non-Hispanic	Female	4	2012
6001	Alameda	2010	1	White. Non-Hispanic	Female	5	2036
6001	Alameda	2010	1	White. Non-Hispanic	Female	6	2114
6001	Alameda	2010	1	White. Non-Hispanic	Female	7	2010
6001	Alameda	2010	1	White. Non-Hispanic	Female	8	2051

4. **[10pts total] [8pts] Forecast the educational demand for California, up to the year of 2060 using the California Educational Attainment and Personal Income data that will provide education information and the California Population Projection by county, age, gender, and ethnicity data that will provide population forecasts up to 2060.** Note that the datasets don't share any fields in common, so a simple join across two columns is impossible and the educational attainment data only spans the years 2008 to 2014, and our projection will need to go from 2010 to 2060. Additionally, you can see that caea dataset shares the year, age, and gender columns with a population projection data, but the year is in a different format and the age is not a numeric age, but rather an age range. This is the field we're going to have to link with the numerical ages in the population projection. As there is no link between the datasets but you can use conditional statements to join age to the age range (00 to 17, 18 to 64, and 65 to 80+).

   [2pts] Additionally, as the demographics table code has been given to you, please describe in words what this table is finding and how you will use this information.

```
CREATE TABLE pop_proj (
  county_code INT,
  county` VARCHAR(45),
  date_year VARCHAR(4),
  race_code INT,
  race TEXT,
  age INT,
  population INT);
```

county_code	county	date_year	race_code	race	gender	age	population
6001	Alameda	2010	1	White. Non-Hispanic	Female	0	2078
6001	Alameda	2010	1	White. Non-Hispanic	Female	1	2038
6001	Alameda	2010	1	White. Non-Hispanic	Female	2	2008
6001	Alameda	2010	1	White. Non-Hispanic	Female	3	2129
6001	Alameda	2010	1	White. Non-Hispanic	Female	4	2012
6001	Alameda	2010	1	White. Non-Hispanic	Female	5	2036
6001	Alameda	2010	1	White. Non-Hispanic	Female	6	2114
6001	Alameda	2010	1	White. Non-Hispanic	Female	7	2010
6001	Alameda	2010	1	White. Non-Hispanic	Female	8	2051

- gender is VARCHAR(6), since the only two options are male and female.

```
CREATE TABLE caea (
   date_year VARCHAR(50) NOT NULL,
   age VARCHAR(50) NOT NULL,
   gender VARCHAR(7) NOT NULL,
   ed_attainment TEXT NOT NULL,
   income TEXT NOT NULL,
   population INT NOT NULL);
```

date_year	age	gender	ed_attainment	income	population
01/01/2008 12:00:00 AM	00 to 17	Male	Children under 15	No Income	0
01/01/2008 12:00:00 AM	00 to 17	Male	No high school diploma	No Income	650889
01/01/2008 12:00:00 AM	00 to 17	Male	No high school diploma	$5.000 to $9.999	30152
01/01/2008 12:00:00 AM	00 to 17	Male	No high school diploma	$10.000 to $14.999	7092
01/01/2008 12:00:00 AM	00 to 17	Male	No high school diploma	$15.000 to $24.999	3974
01/01/2008 12:00:00 AM	00 to 17	Male	No high school diploma	$25.000 to $34.999	2606
01/01/2008 12:00:00 AM	00 to 17	Male	No high school diploma	$35.000 to $49.999	2227
01/01/2008 12:00:00 AM	00 to 17	Male	High school or equivalent	No Income	0
01/01/2008 12:00:00 AM	00 to 17	Male	Some college. less than 4-yr degree	No Income	8664

```
CREATE TABLE demographics AS
SELECT caea.age AS Age,
   ed_attainment AS Education,
   SUM(population) / total_pop_by_age.total_pop AS coefficient
FROM caea
JOIN (SELECT age, SUM(population) AS total_pop
        FROM caea GROUP BY age) AS total_pop_by_age
        ON caea.age = total_pop_by_age.age
GROUP BY Age, Education;
```

age	education	coefficient
00 to 17	Bachelor's degree or higher	0.0015
00 to 17	Children under 15	0.0000
00 to 17	High school or equivalent	0.0117
00 to 17	No high school diploma	0.9774
00 to 17	Some college. less than 4-yr degree	0.0094
18 to 64	Bachelor's degree or higher	0.3032
18 to 64	High school or equivalent	0.2344
18 to 64	No high school diploma	0.1670
18 to 64	Some college. less than 4-yr degree	0.2954

Example of Output

Year	Education	Demand
2010	Bachelor's degree or higher	8496037
2010	Children under 15	0
2010	High school or equivalent	6758925
2010	No high school diploma	13898441
2010	Some college. less than 4-yr degree	8180179
2011	Bachelor's degree or higher	8599364
2011	Children under 15	0
2011	High school or equivalent	6841812
2011	No high school diploma	13958514