Part 1: SQL on Football data

Problem 1. Find all games where Seattle was the home or away team. For each game, return the two teams and scores.

```sql
Select home, homescore, away, awayscore
From Football
Where home = 'Seattle' or away = 'Seattle'
```

Problem 2. Find all games where the home team scored more than 40 points and the away team scored fewer than 10 points. For each game, return the home and away team, their scores, and the prediction for the game; return the result sorted by the home-team score in descending order.

```sql
Select home, homescore, away, awayscore, prediction
From Football
Where homescore > 40 and awayscore < 10
Order By homescore desc
```

Problem 3. Find all teams who lost by more than 20 points as the home team during the first four weeks of some year (i.e., Week < 5). Return the teams in alphabetical order and return each team only once.

```sql
Select Distinct Home
From Football
Where AwayScore - HomeScore > 20 and week < 5
Order By Home
```

Problem 4. Find the average prediction for games where the home team won. Then in a separate query find the average prediction for games where the away team won.

```sql
Select avg(prediction) From Football Where HomeScore > AwayScore
Select avg(prediction) From Football Where HomeScore < AwayScore
```

Problem 5. In each of the three years (1998, 1999, 2000), how many games were there where the home team was predicted to win, i.e., prediction > 0, and the actual point spread (HomeScore - AwayScore) was more than three times the prediction? Your result should have three tuples, each containing the year and number of games.

```sql
Select year, count(*)
From Football
Where prediction > 0 and HomeScore - AwayScore > 3*prediction
Group By year
```

Problem 6. Find all pairs of teams where the two teams played each other in 1998 and 1999 in the same configuration (the same team was home and the same team was away), and in 1998 the home team won while in 1999 the away team won. Return the pairs of teams.
Part 2: SQL on Schoolkids data

Problem 7. Find all students whose goal is Sports but they ranked Sports as fourth in impact on popularity. Return the entire tuple for each such student.

```sql
Select * From Schoolkids Where Goal = 'Sports' and Sports = 4
```

Problem 8. Count how many 4th graders think Looks are more important than Grades. Return just one number. Then run the same query for 5th graders and again for 6th graders, returning one number each time.

```sql
Select count(*) From Schoolkids Where grade = 4 and Looks < Grades
```

Problem 9. For each of the three goals (Grades, Sports, Popular), what is the average age of students who have that goal? Your result should have three tuples, one each with the goal and the average age for that goal.

```sql
Select goal, avg(age) From Schoolkids Group By goal
```

Problem 10. Find all 4th graders in Urban schools where there is no 5th grader in their same school who has the same goal and same ranking of popularity impact factors. Return each student's grade, school, and goal.

```sql
Select grade, school, goal From Schoolkids S1
Where grade = 4 and type = 'Urban' and Not Exists
  (Select * From Schoolkids S2
   Where S2.grade = 5 and S2.school = S1.school and S2.goal = S1.goal
    and S2.grades = S1.grades and S2.sports = S1.sports and S2.looks = S1.looks
    and S2.money = S1.money)
```

Problem 11. Find the oldest students. For each one return their gender, grade, age, and school.

```sql
select grade, age, school
From Schoolkids
Where age >= (Select max(age) From Schoolkids)
```

Part 3: Correlation and causation

Problem 12. Individual solutions

Problem 13. In the Schoolkids data, we might expect there to be some correlation between Gender and Goal. Since both measures are categorical, use the table method shown in class to determine whether there appears to be a correlation. You may use a spreadsheet, SQL, or any other tool you like to create the table. Include the table, then argue based on the table whether or not there appears to be a correlation.
Row percentages are not similar (and, for good measure, neither are column percentages), so there is correlation.

<table>
<thead>
<tr>
<th></th>
<th>boy</th>
<th>row %</th>
<th>girl</th>
<th>row %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grades</td>
<td>117</td>
<td>47.37%</td>
<td>130</td>
<td>52.63%</td>
</tr>
<tr>
<td>Popular</td>
<td>50</td>
<td>35.46%</td>
<td>91</td>
<td>64.54%</td>
</tr>
<tr>
<td>Sports</td>
<td>60</td>
<td>66.67%</td>
<td>30</td>
<td>33.33%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>boy</th>
<th>col %</th>
<th>girl</th>
<th>col %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grades</td>
<td>117</td>
<td>51.54%</td>
<td>130</td>
<td>51.79%</td>
</tr>
<tr>
<td>Popular</td>
<td>50</td>
<td>22.03%</td>
<td>91</td>
<td>36.25%</td>
</tr>
<tr>
<td>Sports</td>
<td>60</td>
<td>26.43%</td>
<td>30</td>
<td>11.95%</td>
</tr>
</tbody>
</table>

**SQL Extra Credit**

**Problem EX1:** On the Football data, find all teams whose average score playing as a home team is lower than its average score playing as an away team. For each team, return the team, its average score in home games, and its average score in away games.

Select F1.Home, avg(HomeScore),
(Select avg(AwayScore) From Football F2 Where F2.Away = F1.home)
From Football F1
Group By Home
Having avg(HomeScore) < (Select avg(AwayScore) From Football F2
Where F2.Away = F1.home)

**Problem EX2:** On the Schoolkids data, revisit Problem 8, but instead of running three separate queries, run one query whose result is a single tuple with six attributes:

- the integer 4
- the number of 4th graders who think Looks are more important than grades
- the integer 5
- the number of 5th graders who think Looks are more important than grades
- the integer 6
- the number of 6th graders who think Looks are more important than grades

Select g1, c1, g2, c2, g3, c3
From (Select grade as g1, count(*) as c1
From Schoolkids Where grade = 4 and Looks < Grades) Four,
(Select grade as g2, count(*) as c2
From Schoolkids Where grade = 5 and Looks < Grades) Five,
(Select grade as g3, count(*) as c3
From Schoolkids Where grade = 6 and Looks < Grades) Six

**Correlation and Causation Extra Credit**
Problem EX3: Find an apparent correlation in the Schoolkids data, other than the possible correlation between Gender and Goal explored in Problem 13. State the correlation and then show why it appears to hold.

Grade and Age appear correlated. Row percentages are not similar (and, for good measure, neither are column percentages).

<table>
<thead>
<tr>
<th>row age, col grade</th>
<th>4</th>
<th>row %</th>
<th>5</th>
<th>row %</th>
<th>6</th>
<th>col %</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>1</td>
<td>100.00%</td>
<td></td>
<td>0.00%</td>
<td></td>
<td>0.00%</td>
</tr>
<tr>
<td>9</td>
<td>97</td>
<td>97.00%</td>
<td>3</td>
<td>3.00%</td>
<td></td>
<td>0.00%</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>15.15%</td>
<td>112</td>
<td>84.85%</td>
<td></td>
<td>0.00%</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0.53%</td>
<td>58</td>
<td>30.69%</td>
<td>130</td>
<td>68.78%</td>
</tr>
<tr>
<td>12</td>
<td>0.00%</td>
<td>3</td>
<td>5.77%</td>
<td>49</td>
<td>94.23%</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.00%</td>
<td></td>
<td>0.00%</td>
<td>4</td>
<td>100.00%</td>
<td></td>
</tr>
</tbody>
</table>

Problem EX4: Find an apparent correlation in the Football data. State the correlation and then show why it appears to hold.

Correlation between team and HomeScore, i.e., some teams appear to be better than others! Bar graph of teams and average HomeScore:

Also apparent correlation between Prediction and (HomeScore - AwayScore), as shown in the scatterplot for Problem 13 on Assignment #1.