

One assumption I've often made about regions, and which I would like to test using Regions Project data, is the following : In general, the more influential we consider a place, the more we will like it. I think this stems from a basic desire to see ourselves, and the places we are from, in a positive light. Is this true? Testing this hypothesis required a number of steps and aggregations of data, which are detailed below.

First, a quick introduction to how the data is stored. Each map contains a number of regions. Each region, in turn, contains a number of boxes. Each map also contains a number of boxes which are marked as 'influential'.

**Step 1 : Determine which regions contain an 'influential' place.**

This query creates a table containing the regionIDs of all regions which contain an influential box from the map which they are a member of.

```
create table tmp_influentialregion (
  regionID INT NOT NULL PRIMARY KEY
)
SELECT DISTINCT
  r.regionID
FROM influencedata id
  inner join mapdata md using (boxNumber)
  inner join region r using (regionID)
  inner join map m on r.mapID = m.mapID and id.mapID =
m.mapID;
```

**Step 2 : Add 'influential' data to main 'region' table.**

This is a violation of strict data normalization, but simplifies subsequent queries by eliminating the need to left join to the tmp\_influentialregion table. First alter the 'region' table to add an 'isInfluential' enumeration, which is initially set to 'U' (unknown). Next update the value so any region which can join to tmp\_influentialregion has its 'isInfluential' value set to 'Y'. Finally, set all remaining 'U' values to 'N'. Any regions which are created after this point will have 'U' as their isInfluential value, so they will be distinguishable from the 'Y' and 'N' values we have just calculated.

```
ALTER TABLE region ADD isInfluential ENUM('Y','N','U') DEFAULT
'U';

UPDATE region, tmp_influentialregion
SET region.isInfluential = 'Y'
WHERE
  region.regionID = tmp_influentialregion.regionID;

UPDATE region
SET isInfluential = 'N'
WHERE isInfluential = 'U';
```

**Step 3 : Compute average impression, time spent, and well-known values for each standard region, grouping by 'isInfluential'. Save in summary table.**

Each region is part of a 'standard region' group. These are groups of regions which have names differing in minor ways, and which are judged to be referring to the same region. Comparing standard regions, rather than comparing regions strictly on the

name submitted by users, allows more meaningful comparison.

The query excludes regions which are marked as problematic (r.problem = 'Y') or which are part of the standard regions 'Not Standardized' (not yet reviewed, standardRegionID 1) and 'Not Standardizable' (judged to not be similar to any other regional names submitted, standardRegionID 2).

```
CREATE TABLE tmp_influentialimpression (
    standardRegionID INT NOT NULL,
    isInfluential ENUM('Y','N','U') NOT NULL,
    regions_count INT NOT NULL,
    impression_avg DECIMAL(6,4),
    wellKnown_avg DECIMAL(6,4),
    timeSpent_avg DECIMAL(6,4),
    PRIMARY KEY (standardRegionID,isInfluential)
)
SELECT
    r.standardRegionID,
    r.isInfluential,
    COUNT(r.regionID) as regions_count,
    AVG(r.impression) as impression_avg,
    AVG(r.wellKnown) as wellKnown_avg,
    AVG(r.timeSpent) as timeSpent_avg
FROM region r
WHERE
    r.problem = 'N'
    AND r.standardRegionID NOT IN (1,2)
GROUP BY r.standardRegionID, r.isInfluential;
```

**Step 4 : Query new summary table to look for correlations between average 'impression' value and the presence/absence of an 'influential' box in the region.**

This query compares the average impression of standard region when it does and does not contain an 'influential' box. It tallies the number of times the average impression is greater when the region contains or does not contain an 'influential' box.

The query uses a 'self-join' to join two instances of the 'tmp\_influentialimpression' table, allowing direct comparison of influential and non-influential regions within the same standard region group.

```
SELECT
    SUM(IF(iiy.impression_avg = iin.impression_avg,1,0)) as
    'equal',
    SUM(IF(iiy.impression_avg > iin.impression_avg,1,0)) as
    'yes greater',
    SUM(IF(iiy.impression_avg < iin.impression_avg,1,0)) as
    'no greater' ,
    SUM(IF(iiy.impression_avg IS NULL OR iin.impression_avg
    IS NULL,1,0)) as 'no comparison'

FROM standardregion sr
    LEFT JOIN tmp_influentialimpression iiy ON
sr.standardRegionID = iiy.standardRegionID AND
iiy.isInfluential = 'Y'
    LEFT JOIN tmp_influentialimpression iin ON
sr.standardRegionID = iin.standardRegionID AND
iin.isInfluential = 'N'
#exclude 'not standardized', 'not standardizable', and 'ocean'
WHERE sr.standardRegionID NOT IN (1,2,19)
```

Running this final query yields :

<b>equal</b>	<b>yes greater</b>	<b>no greater</b>	<b>No comparison</b>
2	25	5	3

In percentage terms, this is :

<b>equal</b>	<b>yes greater</b>	<b>no greater</b>	
5.71%	71.42%	14.29%	8.57%

-- in non-aggregated form...

SQL result

Host: localhost

Database: regions

Generation Time: Feb 27, 2005 at 09:52 PM

Generated by: phpMyAdmin 2.6.0-pl2 / MySQL 4.0.22-standard

SQL-query: SELECT sr.name, IFNULL(iiy.impression\_avg,'No Data') as 'Avg Impression When Influential', IFNULL(iin.impression\_avg,'No Data') as 'Avg Impression When Not Influential', IF(iiy.impression\_avg = iin.impression\_avg,1,0) as 'Equal', IF(iiy.impression\_avg > iin.impression\_avg,1,0) as 'Avg Impression Greater When Influential', IF(iiy.impression\_avg < iin.impression\_avg,1,0) as 'Avg Impression Greater When Not Influential' , IF(iiy.impression\_avg IS NULL OR iin.impression\_avg IS NULL,1,0) as 'No Comparison Possible' FROM standardregion sr LEFT JOIN tmp\_influentialimpression iiy ON sr.standardRegionID = iiy.standardRegionID AND iiy.isInfluential = 'Y' LEFT JOIN tmp\_influentialimpression iin ON sr.standardRegionID = iin.standardRegionID AND iin.isInfluential = 'N' where sr.standardRegionID NOT IN (1,2,19) order by sr.name LIMIT 0, 100;

Rows: 36

name	Avg Impression When Influential			Avg Impression When Not Influential		
	Equal	Avg Impression Greater	No Comparison Possible	Equal	Avg Impression Greater	No Comparison Possible
Alaska	2.0000	1.7500	0	1	0	0
Appalachia	2.0000	0.0000	0	1	0	0
California	2.0000	0.0000	0	1	0	0
California - Southern	0	-1.3333	-2.0000	0	1	0
Central	0.7500	1.5000	0	0	1	0
Central - South	0	1.0000	1.0000	1	0	0
Deep South	-1.2500	-0.1667	0	0	1	0
Desert	No Data	-0.7500	0	0	0	1
East	0.7273	1.4444	0	0	1	0
East Coast	1.0000	1.4286	0	0	1	0
Florida/Caribbean	0	1.4000	1.0000	0	1	0
Great Lakes	1.6250	1.1667	0	1	0	0
Gulf	3.0000	1.0000	0	1	0	0
Hawaii	3.2500	2.3158	0	1	0	0
Industrial	-2.0000	-1.0000	0	0	1	0
Mid Atlantic	1.0000	0.5000	0	1	0	0
Middle	1.0000	1.0000	1	0	0	0
Midwest	1.4151	0.8421	0	1	0	0
Mountain	2.3750	1.4286	0	1	0	0
New England	2.0667	1.5882	0	1	0	0

North Central	2.5000	1.5000	0	1	0	0
Northeast	2.4545	0.8333	0	1	0	0
Northern	0.8000	-0.2500	0	1	0	0
Northwest	2.8485	1.8750	0	1	0	0
Ocean	No Data	No Data	0	0	0	1
Pacific Coast	1.8500	1.0000	0	1	0	0
Plains	0.6667	0.3333	0	1	0	0
Plains - North	1.5000	1.0000	0	1	0	0
Rocky Mountain		3.4167	1.5000	0	1	0
	0					
South	0.2424	-0.0714	0	1	0	0
South - Mid	No Data	0.0000	0	0	0	1
Southeast	0.5833	0.5000	0	1	0	0
Southwest	2.0286	0.9048	0	1	0	0
Texas	0.0000	-1.8182	0	1	0	0
Upper Midwest		3.3750	No Data	0	0	0
	1					
West	2.6897	2.3750	0	1	0	0

#### Discussion :

This line of investigation seems to reveal a clear association between regions which contain an 'influential' box and regions which have a higher average 'impression' score. This would suggest that people tend to like the places they see as influential.