Nolan Winkler

npw@uchicago.edu 1-(773)-767-6117

8/2/2016

Part 1: SQL

1. Select a list of all licenses are not in the city of Chicago.

SELECT * FROM licenses WHERE City NOT LIKE 'Chicago' LIMIT 10;

2. Provide a count of inspections by inspection type.

Using SELECT inspection_type, count(*) FROM inspections GROUP BY inspection_type LIMIT 10;

inspection_type	count(*)
Canvass	3368
Canvass Re-Inspection	590
Complaint	329
Complaint Re-Inspection	167
Consultation	13
License	632
License Re-Inspection	158
Recent Inspection	8
Short Form Complaint	229
Suspected Food Poisoning	10

3. How many zip codes ONLY have 1 license?

266 according to

SELECT COUNT(*) FROM (SELECT zip, count(*) AS num FROM licenses GROUP BY zip) WHERE num = 1;

4. List all the inspections in the zip code with the most inspections. Do not hardcode any zip codes into your code.

SELECT * FROM inspections where zip IN (SELECT zip FROM inspections GROUP BY zip ORDER BY count(*) DESC LIMIT 1);

5. SELECT l.license_num, substr(risk, 6, 1) as risk_value FROM inspections i INNER JOIN licenses l ON i.license_num = l.license_num WHERE dba_name LIKE 'Subway' OR aka_name like 'Subway' OR business_name like 'Subway';

6. SELECT risk, (appropriate regexp on address) as street_direction, count(*) FROM inspections WHERE substr(risk, 6, 1) == '3' GROUP BY street_direction LIMIT 10;

7. SELECT dba_name, inspection_id, inspection_date, inspection_type, results from inspections where license_num IN (SELECT license_num FROM (SELECT i.license_num, count(*) as num FROM inspections i inner join licenses l on i.license_num = l.license_num GROUP BY i.license_num) WHERE num > 1);

dba_name	inspection_id	inspection_date	inspection_type	results
LE PITA	1464668	20141224	Canvass Re-Inspection	Pass
FLYING WOK INC.	1515372	20141223	Canvass Re-Inspection	Pass w/ Conditions
FLYING WOK INC.	1515378	20141223	Canvass	Out of Business
LE PITA	1464646	20141217	Canvass	Fail

LAVASH	1199662	20141216	Canvass Re-Inspection	Pass
FLYING WOK INC.	1513104	20141216	Canvass	Fail
BERNIE'S RESTAURANT	1513126	20141216	Canvass Re-Inspection	Pass
ONESTI PIZZERIA INC	1513127	20141216	Complaint Re-Inspection	Pass
TAJ MAHAL	1513027	20141215	Canvass Re-Inspection	Pass w/ Conditions
THE FLAMING POT INC.	1512950	20141212	Complaint Re-Inspection	Pass

8. SELECT * from inspections where dba_name like 'Burger King' or aka_name like 'Burger King' ORDER BY dba_name;

Part 2: Research

1) **Consumer Expenditure Survey - http://www.bls.gov/cex/tables.htm**

-> The CES is the only Federal survey to provide information on the complete range of consumers' expenditures and incomes, as well as the characteristics of those consumers.

-> This data would be extremely valuable as it allows us to link the address, age, and gender of the person to summary statistics on peoples' incomes in their geographic region, for their age, and for their gender.

-> Each person's address could be used to link them to their geographic region, state, and metropolitan area if applicable. Using that information, we can then get out the predicted average income for people in their age range and gender. We can then use this information as essentially labeled data for our first-attempt supervised learning approach using the BLS' averages as our responses.

2) National Health Interview Survey - http://www.cdc.gov/nchs/nhis/shs/tables.htm

-> The annual Summary Health Statistics tables summarize data from the National Health Interview Survey (NHIS), a multipurpose health survey conducted by the Centers for Disease Control and Prevention's (CDC) National Center for Health Statistics (NCHS). These tables provide national estimates for a broad range of health measures for the U.S. civilian noninstitutionalized population.

-> This data would give us predicted information about the person's health situation via their address, age, and gender by looking at summary statistics for people in that same category.

-> Each person's address could be used to link them to their geographic region, state, and metropolitan area if applicable. Using that information, we can then get out the predicted average income for people in their age range and gender. We can then use this information as essentially labeled data for our first-attempt supervised learning approach using the BLS' averages as our responses.

3) Education section of the US Census - http://www.census.gov/library/publications/2011/compendia/statab/131ed/education.html

http://www.census.gov/hhes/socdemo/education/data/cps/2015/tables.html

-> The Education section presents data primarily concerning formal education as a whole, at various levels, and for public and private schools. Data shown relate to the school-age

population and school enrollment, educational attainment, education personnel, and financial aspects of education.

-> This data can be used to give us some semblance of the educational status of our individuals based on their geographic location. In particular, we can very roughly estimate their educational attainment based on this.

-> I would link and use this data similarly to the two above examples.

4) Demographic race data - http://factfinder.census.gov/faces/nav/jsf/pages/community_facts.xhtml

-> A breakdown of racial demographics by ZIP code from the US census

-> I would use this data to have better predictions for one's race class in our model.

-> I could not find a simple table that does this for each ZIP code, so we would have to download it for each ZIP code, create a table with the ZIPs and their corresponding information, and finally link those ZIPs to people via their addresses.

Data sets I can't find but think would be useful for this project:

1) Medical records/summaries for each person in the US - includes height, disabilities, and weight.

-> I think this data would be useful because height seems to have been shown to be correlated with income, I imagine weight might be as well, and clearly if somebody is in a bad state of health or has a disabling medical condition, they may be earning less. This data in particular could be very helpful for creating clusters and being able to do hierarchical models or regressions based on decision trees. As I imagine there would be a fairly small subset of the data that would be most useful (e.g. the factors listed above), as long as extracting that data is not too difficult, then linking it to the person file should be very straightforward pure joining on person name.

2) Educational summaries for each person in the US - number of school years completed, degrees obtained in what fields from which schools

-> This data would be very useful as more education is sure to correlate with more earnings, and knowing which specific schools and programs were attended by individuals would once again be incredibly useful for clustering and separating the data.

3) Genealogies of each person in the US - simple linkage of people to their ancestry.

-> This data would be extremely useful as wealth and income tend to accumulate in families and any information that it is useful to know about a person, it is similarly useful to know about their families, as their lineage could be a great indicator of their earnings.

Part 3: Data Standardization

1) I would first convert each file into a standardized file type such as .csv, which we should be able to export to from almost any program. Next, I would decide upon the information that we would like to have as our columns for the national database and for every file, purge each column that we will not be including. Similarly, many of these files would not have columns that we most definitely would want, so those would have to be added. Then, deciding upon a universal primary key such as a combination of the state that the row is coming from and its primary key in that state's file, I would combine all of these files together into one big .csv which we can then take into our favorite data analysis tool of choice.
2) To cleanse and standardize the data, I would first go through and remove any duplicates within each individual state .csv before merging them all together, as an easy way to detect and remove duplicates in one file might conflict with what we have to do in another file. Then, I would check for missing values in each column and depending on the column type, either mark it explictly as missing or try to impute an appropriate value from some subset of the rest of the data. For example, if SSN is one of the columns that we want, we certainly would have to just leave that blank, but if, say, some data such as predicted income were included, we could guess that based upon the mean of similar rows. PanDas and NumPy/SciPy would essentially be my tools of choice for working with this as it has very nice detecting and filling in missing value functions as well as tools for imputing data.
3) Technical challenges that could be faced would be centered around the quantity of the data and being able to nicely inspect it when it is so large. If there is some proprietary software that is the only way we have to access a particular state's data and that software does not allow us to nicely export it, that would be a problem.
4) Data quality would be ensured by the purging of duplicates using tools such as Dedupe and building a system that would help us to update the data as new publications are made by the individual state databases. In terms of storage and management, using AWS would be the way to go.
5) Other factors that would need to be considered would be what exactly we are using this list of nursing data for. In particular, it is hard to not be vague about what columns we would want to universally have and whether or not we should, say, keep a column that, say, half of the states use and half don't so we would have to either leave blank or use

some estimate for those states if we do not know what purpose the data is being used for. Additionally, how current we want to keep the data, how we are going to be able to link it up with other data that we might want to, and whether we want to archive snapshots of it or overwrite it could all be important considerations. Essentially, there are lots of questions around management of the data that could only be answered by knowing for what potential models or projects we would be using this data.