Cleaning data

Julia Coccia

Data source 1

Industry employment by state

<http://www.bebr.ufl.edu/data/26/state/12000-state-florida>

This data had employment data for the hospitality industry from 1990-2016. It had other states as well as Florida which can be used as a quick comparison between states. I did a lot of clean up in excel before putting the set into python.

Note from data source: ((SA) Seasonally adjusted numbers take into account seasonal hiring/layoff patterns that go along with winter and summer holidays and student summer hires. Total persons on payroll full or part time for any part of the pay period which includes the 12th day of the month. Temporary and intermittent employees, workers on paid sick leave or holiday, or who work during part of the specified pay period, including striking workers, are included. Persons on payroll of more than one establishment are counted in each. Data exclude military, proprietors, self-employed, unpaid family or volunteers, farm and domestic workers. Those on layoff, leave without pay, strike, or not yet reported for work, the entire pay period, are not counted.)

SOURCE: Bureau of Labor Statistics (BLS)

In excel:

1: new headers

2: fill in data into new headers

3: multiply numbers by 1000 (the data was represented by 1000 people)

4: Split date into (month and year)

5.Create a new work book with cleaned data to be used in project called tourism industry employment from 1990-2016

(may add two columns, 1 state number (12 for Florida), 2 state name (Florida), I may add these to connect the employment to other tables for my data set.

Saved as toursimemployment

Data Source 2:

<http://www.floridatransportationindicators.org/index.php?chart=9b&view=detail>

this data has the number of passengers each Florida airport has on a monthly basis

in excel:

1.turn the airport’s into columns and the passenger’s columns into data for those columns. (for this I did a pivot table with rows having the date, columns having the airport names.

2. Separated Month and year from date.

Saved as

Missing Values:

Jacksonville airport had a missing value for Jan,2007 I found the average passengers was 227156 so I replaced the missing with this number

Orlando airport was missing information for December 2001 found out Orlando Sandford airport was listed twice with 2 different passenger amounts looked at the data for Orlando and Sanford noticed Orlando was always significantly higher so I changed the higher Sanford airport to be Orlando airport.

For this data set I left the original data set in case that is the better way to look at it, but I also created a new data set called airport information by passengers for Florida. This new data sets breaks the original data set out by airport and date. I did this because I believe I may be looking at individual cities for impact as I continue my analysis.

Saved as airportdata1

And airportdata2

Data Source 3

<http://www.floridatransportationindicators.org/index.php?chart=11a&view=detail>

Quartly visitors

Data clean up.

Split up quarter/year into quarter and year

Added state name and state id

Saved as floridavisitorsquarterly

date source #4

<http://www.floridatransportationindicators.org/index.php?chart=2b&view=detail>

estimated Florida population by year

delete first row

Importing and joining

To start I am importing my data into SAS

Import codes

**Import airport1**

**DATA** WORK.airportdata1;

LENGTH

date $ **9**

'Ft.Lauderdale\_Hollywood\_Airport'n **8**

Jacksonville\_Airport **8**

Miami\_Airport **8**

'Orlando \_International\_Airport'n **8**

Orlando\_Sanford\_Airport **8**

Palm\_Beach\_Airport **8**

Sarasota\_Bradenton\_Airport **8**

Southwest\_Florida\_Airport **8**

Tampa\_Airport **8**

Month $ **3**

Year $ **4**

Month\_number **8**

Month\_name $ **9** ;

FORMAT

date $CHAR9.

'Ft.Lauderdale\_Hollywood\_Airport'n BEST12.

Jacksonville\_Airport BEST12.

Miami\_Airport BEST12.

'Orlando \_International\_Airport'n BEST12.

Orlando\_Sanford\_Airport BEST12.

Palm\_Beach\_Airport BEST12.

Sarasota\_Bradenton\_Airport BEST12.

Southwest\_Florida\_Airport BEST12.

Tampa\_Airport BEST12.

Month $CHAR3.

Year $CHAR4.

Month\_number BEST12.

Month\_name $CHAR9. ;

INFORMAT

date $CHAR9.

'Ft.Lauderdale\_Hollywood\_Airport'n BEST12.

Jacksonville\_Airport BEST12.

Miami\_Airport BEST12.

'Orlando \_International\_Airport'n BEST12.

Orlando\_Sanford\_Airport BEST12.

Palm\_Beach\_Airport BEST12.

Sarasota\_Bradenton\_Airport BEST12.

Southwest\_Florida\_Airport BEST12.

Tampa\_Airport BEST12.

Month $CHAR3.

Year $CHAR4.

Month\_number BEST12.

Month\_name $CHAR9. ;

INFILE '/saswork/SAS\_work441A00015CE9\_odaws04-prod-us/#LN00010'

LRECL=**95**

ENCODING="UTF-8"

TERMSTR=CRLF

DLM='7F'x

MISSOVER

DSD ;

INPUT

date : $CHAR9.

'Ft.Lauderdale\_Hollywood\_Airport'n : BEST32.

Jacksonville\_Airport : BEST32.

Miami\_Airport : BEST32.

'Orlando \_International\_Airport'n : BEST32.

Orlando\_Sanford\_Airport : BEST32.

Palm\_Beach\_Airport : BEST32.

Sarasota\_Bradenton\_Airport : BEST32.

Southwest\_Florida\_Airport : BEST32.

Tampa\_Airport : BEST32.

Month : $CHAR3.

Year : $CHAR4.

Month\_number : BEST32.

Month\_name : $CHAR9. ;

**RUN**;

libname CIS '/home/juliacoccia/my\_courses/cis/';

**data** CIS.airport1;

set WORK.airportdata1;

**run**;

**import tourism employment**

-------------------------------------------------------------------- \*/

**DATA** WORK.toursimemployment;

LENGTH

date **8**

alabama **8**

alaska **8**

arizona **8**

arkansas **8**

california **8**

colorado **8**

connecticut **8**

delaware **8**

district\_of\_columbia **8**

florida **8**

georgia **8**

hawaii **8**

idaho **8**

illinois **8**

indiana **8**

iowa **8**

kansas **8**

kentucky **8**

louisiana **8**

maine **8**

maryland **8**

massachusetts **8**

michigan **8**

minnesota **8**

mississippi **8**

missouri **8**

montana **8**

nebraska **8**

nevada **8**

new\_hampshire **8**

new\_jersey **8**

new\_mexico **8**

new\_york **8**

north\_carolina **8**

north\_dakota **8**

ohio **8**

oklahoma **8**

oregon **8**

pennsylvania **8**

rhode\_island **8**

south\_carolina **8**

south\_dakota **8**

tennessee **8**

texas **8**

utah **8**

vermont **8**

virginia **8**

washington **8**

west\_virginia **8**

wisconsin **8**

Month **8**

Year **8**

Month\_name $ **9**

Florida\_name $ **7**

State\_id **8** ;

FORMAT

date DATE9.

alabama F12.2

alaska F12.2

arizona F12.2

arkansas F12.2

california F12.2

colorado F12.2

connecticut F12.2

delaware F12.2

district\_of\_columbia F12.2

florida F12.2

georgia F12.2

hawaii F12.2

idaho F12.2

illinois F12.2

indiana F12.2

iowa F12.2

kansas F12.2

kentucky F12.2

louisiana F12.2

maine F12.2

maryland F12.2

massachusetts F12.2

michigan F12.2

minnesota F12.2

mississippi F12.2

missouri F12.2

montana F12.2

nebraska F12.2

nevada F12.2

new\_hampshire F12.2

new\_jersey F12.2

new\_mexico F12.2

new\_york F12.2

north\_carolina F12.2

north\_dakota F12.2

ohio F12.2

oklahoma F12.2

oregon F12.2

pennsylvania F12.2

rhode\_island F12.2

south\_carolina F12.2

south\_dakota F12.2

tennessee F12.2

texas F12.2

utah F12.2

vermont F12.2

virginia F12.2

washington F12.2

west\_virginia F12.2

wisconsin F12.2

Month F12.

Year F12.

Month\_name $CHAR9.

Florida\_name $CHAR7.

State\_id BEST12. ;

INFORMAT

date DATE9.

alabama BEST12.

alaska BEST12.

arizona BEST12.

arkansas BEST12.

california BEST12.

colorado BEST12.

connecticut BEST12.

delaware BEST12.

district\_of\_columbia BEST12.

florida BEST12.

georgia BEST12.

hawaii BEST12.

idaho BEST12.

illinois BEST12.

indiana BEST12.

iowa BEST12.

kansas BEST12.

kentucky BEST12.

louisiana BEST12.

maine BEST12.

maryland BEST12.

massachusetts BEST12.

michigan BEST12.

minnesota BEST12.

mississippi BEST12.

missouri BEST12.

montana BEST12.

nebraska BEST12.

nevada BEST12.

new\_hampshire BEST12.

new\_jersey BEST12.

new\_mexico BEST12.

new\_york BEST12.

north\_carolina BEST12.

north\_dakota BEST12.

ohio BEST12.

oklahoma BEST12.

oregon BEST12.

pennsylvania BEST12.

rhode\_island BEST12.

south\_carolina BEST12.

south\_dakota BEST12.

tennessee BEST12.

texas BEST12.

utah BEST12.

vermont BEST12.

virginia BEST12.

washington BEST12.

west\_virginia BEST12.

wisconsin BEST12.

Month BEST12.

Year BEST12.

Month\_name $CHAR9.

Florida\_name $CHAR7.

State\_id BEST12. ;

INFILE '/saswork/SAS\_work441A00015CE9\_odaws04-prod-us/#LN00027'

LRECL=**372**

ENCODING="UTF-8"

TERMSTR=CRLF

DLM='7F'x

MISSOVER

DSD ;

INPUT

date : BEST32.

alabama : BEST32.

alaska : BEST32.

arizona : BEST32.

arkansas : BEST32.

california : BEST32.

colorado : BEST32.

connecticut : BEST32.

delaware : BEST32.

district\_of\_columbia : BEST32.

florida : BEST32.

georgia : BEST32.

hawaii : BEST32.

idaho : BEST32.

illinois : BEST32.

indiana : BEST32.

iowa : BEST32.

kansas : BEST32.

kentucky : BEST32.

louisiana : BEST32.

maine : BEST32.

maryland : BEST32.

massachusetts : BEST32.

michigan : BEST32.

minnesota : BEST32.

mississippi : BEST32.

missouri : BEST32.

montana : BEST32.

nebraska : BEST32.

nevada : BEST32.

new\_hampshire : BEST32.

new\_jersey : BEST32.

new\_mexico : BEST32.

new\_york : BEST32.

north\_carolina : BEST32.

north\_dakota : BEST32.

ohio : BEST32.

oklahoma : BEST32.

oregon : BEST32.

pennsylvania : BEST32.

rhode\_island : BEST32.

south\_carolina : BEST32.

south\_dakota : BEST32.

tennessee : BEST32.

texas : BEST32.

utah : BEST32.

vermont : BEST32.

virginia : BEST32.

washington : BEST32.

west\_virginia : BEST32.

wisconsin : BEST32.

Month : BEST32.

Year : BEST32.

Month\_name : $CHAR9.

Florida\_name : $CHAR7.

State\_id : BEST32. ;

**RUN**;

libname CIS '/home/juliacoccia/my\_courses/cis/';

**data** CIS.employment;

set WORK.toursimemployment;

**run**;

import visitors

**DATA** WORK.FLORIDAVISITORSQUARTERLY\_0000;

LENGTH

Quarter $ **7**

Visitors **8**

year **8**

quarter\_0001 $ **2**

State $ **7**

State\_number **8** ;

LABEL

quarter\_0001 = "quarter" ;

FORMAT

Quarter $CHAR7.

Visitors BEST12.

year BEST12.

quarter\_0001 $CHAR2.

State $CHAR7.

State\_number BEST12. ;

INFORMAT

Quarter $CHAR7.

Visitors BEST12.

year BEST12.

quarter\_0001 $CHAR2.

State $CHAR7.

State\_number BEST12. ;

INFILE '/saswork/SAS\_workC1860000DB9F\_odaws04-prod-us/#LN00046'

LRECL=**35**

ENCODING="UTF-8"

TERMSTR=CRLF

DLM='7F'x

MISSOVER

DSD ;

INPUT

Quarter : $CHAR7.

Visitors : BEST32.

year : BEST4.

quarter\_0001 : $CHAR2.

State : $CHAR7.

State\_number : BEST32. ;

**RUN**;

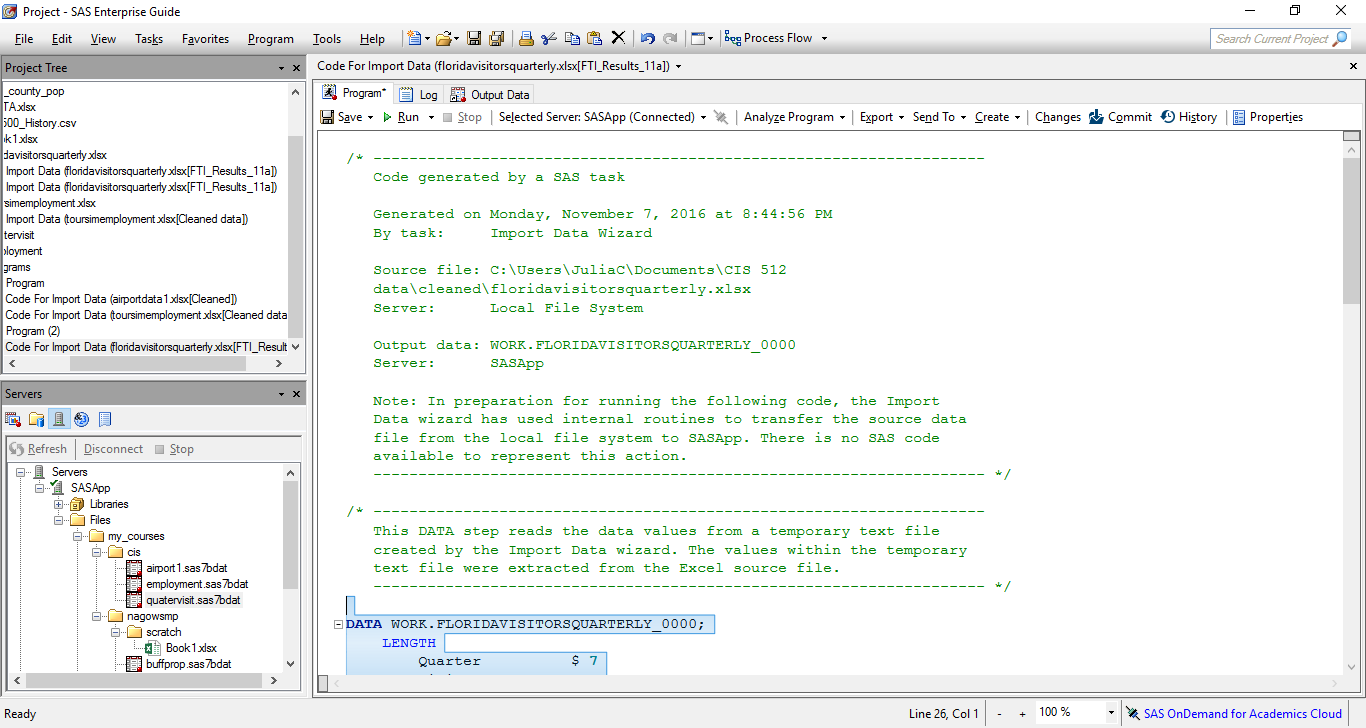
libname CIS '/home/juliacoccia/my\_courses/cis/';

**data** CIS.quatervisit;

set WORK.FLORIDAVISITORSQUARTERLY\_0000;

**run**;

**Using the Libname statement I was able to create a library here is a screen shot of my data stored in my newly created library:**



**program to join two data sets as well as to continue to clean data**

libname CIS '/home/juliacoccia/my\_courses/cis/';

**proc** **contents** data=CIS.employment;

**run**;

**proc** **contents** data=CIS.quatervisit;

**run**;

**proc** **sql**;

create table employ as

select date,florida as employees, Month,Month\_name,year, case when Month between **1** and **3** then 'Q1'

when Month between **4** and **6** then 'Q2'

when Month between **7** and **9** then 'Q3'

when Month between **10** and **12** then 'Q4'

end as quarter

from CIS.employment;

**quit**;

**proc** **sql**;

create table employee as

select sum(employees)as employee\_qtr, year,quarter

from employ

group by quarter, year;

**quit**;

**proc** **contents** data=employee;

**run**;

**proc** **sql**;

create table CIS.Visit as

Select a.employee\_qtr,a.year,a.quarter,Visitors,b.Quarter

From employee as a

left outer join cis.quatervisit as b on

a.quarter=b.quarter\_0001

and a.year=b.year;

**quit**;

I have attached an excel of the data joined to my GitHub called Visit. This data is by quarterly totals; it has a lot of missing data but that is because one data set had less years of data. When I am doing an analysis between these two sets ill only look at years where they both have data