Paper 1940-2016

Case Study: SAS® Visual Analytics Dashboard for Pollution Analysis

Viraj Kumbhakarna, MUFG Union Bank, San Francisco, CA

Baskar Anjappan, MUFG Union Bank, San Francisco, CA

ABSTRACT

This paper is a case study to explore the analytical and reporting capabilities of SAS Visual Analytics to perform data exploration, determine order patterns and trends, and create data visualizations to generate extensive dashboard reports using the open source pollution data available from the US Environmental States Environment Protection Agency.

Data collection agencies report their data to EPA via the system called Air Quality System (AQS) and it makes available several types of aggregate (summary) datasets such as daily and annual pollutant summaries in CSV format for public use. We intend to demonstrate SAS Visual Analytics capabilities by using pollution data to create visualizations which compare Air Quality Index (AQI) values for multiple pollutants by location and time period, generate a time series plot by location and time period, compare 8-hour ozone "exceedances" from this year with previous years and perform other such analysis.

Easy to use SAS Visual Analytics web-based interface will be leveraged to explore patterns in the pollutant data to obtain insightful information. SAS Visual data builder will be used to summarize data, join datasets and enhance the predictive power within the data. SAS Visual Analytics explorer will be used to explore data, to create calculated data items, aggregated measures, define geography items. Visualizations such as chats, bar graphs, geo maps, tree maps, correlation matrices and other graphs will be created to graphically visualize pollutant information contaminating the environment; hierarchies will be derived from date, time items and across geographies to allow rolling up the data. Various reports will be designed for pollution analysis and viewing reports on a mobile device such as an iPad will be also explored. In conclusion, in this paper will attempt to demonstrate use of SAS Visual Analytics to determine impact of pollution on the environment over time using various visualizations and graphs.

INTRODUCTION

The purpose of the paper is to give brief overview about capabilities of the SAS Visual Analytics dashboard. Many users experienced in Base SAS programming and other reporting procedures in SAS can utilize the capabilities of the SAS Visual Analytics to create reports and data visualizations. SAS Visual Analytics tool makes a user's life easy by allowing a powerful way to analyze the datasets to create graphs and reports using various SAS Visual Analytics functions.

SAS Visual Analytics provides a complete platform for analytics visualization, enabling you to identify patterns and relationships in data that weren't initially evident. Interactive, self-service BI and reporting capabilities are combined with out-of-the-box advanced analytics so everyone can discover insights from any size and type of data, including text. Users of all skill levels can visually explore data on their own while tapping into powerful in-memory technologies for faster analytic computations and discoveries. It's an easy-to-use, self-service environment that can scale on an enterprise wide level.

In this paper we explore via a case study how to utilize SAS Visual Analytics to data, perform data transformation and cleansing and explore data to analyze trends and create visualizations on third party pollution data.

ENVIRONMENTAL PROTECTION AGENCY (EPA) DATA SOURCE

During the course of this case study the open source air pollution data was obtained from the US Environmental Protection Agency (EPA) website for creating dashboards using SAS Visual Analytics. The United States Environmental Protection Agency (EPA) is an agency of the U.S. federal government which was created for the purpose of protecting human health and the environment by writing and enforcing regulations based on laws passed by Congress. The EPA has its headquarters in Washington, D.C., regional offices for each of the agency's ten regions, and 27 laboratories. The agency conducts environmental assessment, research, and education.

EPA conducts air research to protect human health and the environment in support of the Clean Air Act and National Ambient Air Quality Standards. The AirData website (http://www.epa.gov/airdata/ad_basic.html) gives users access to air quality data collected at outdoor monitors across the United States, Puerto Rico, and the U. S. Virgin Islands. The data comes primarily from the AQS (Air Quality System) database. AirData lets users display and download monitored hourly, daily, and annual concentration data, AQI data, and speciated particle pollution data

SOURCE AQS OZONE MONITOR AIRDATA FROM EPA

For the purpose of this case study we will focus on a subset of the AQS data - the Ozone Monitor data from the Air Quality System (AQS) database. The AQS ozone data across years is available for download from the EPA website from the web link (http://www.epa.gov/heasd/research/cdc.html). The AQS Ozone files contain the maximum 8-hour averages during a 24 hour period of ambient ozone concentration in parts per million (ppm), by state, county and/or MSA, as measured by EPA's Federal reference method (FRM) AQS monitors. The measured concentration data obtained from EPA's AQS monitors is stored in a database along with other attributes, e.g., sample date, state, county, MSA, etc. An EPA computer program is used to extract the concentration data and associated attributes from the database and incorporate it into a *.csv format file.

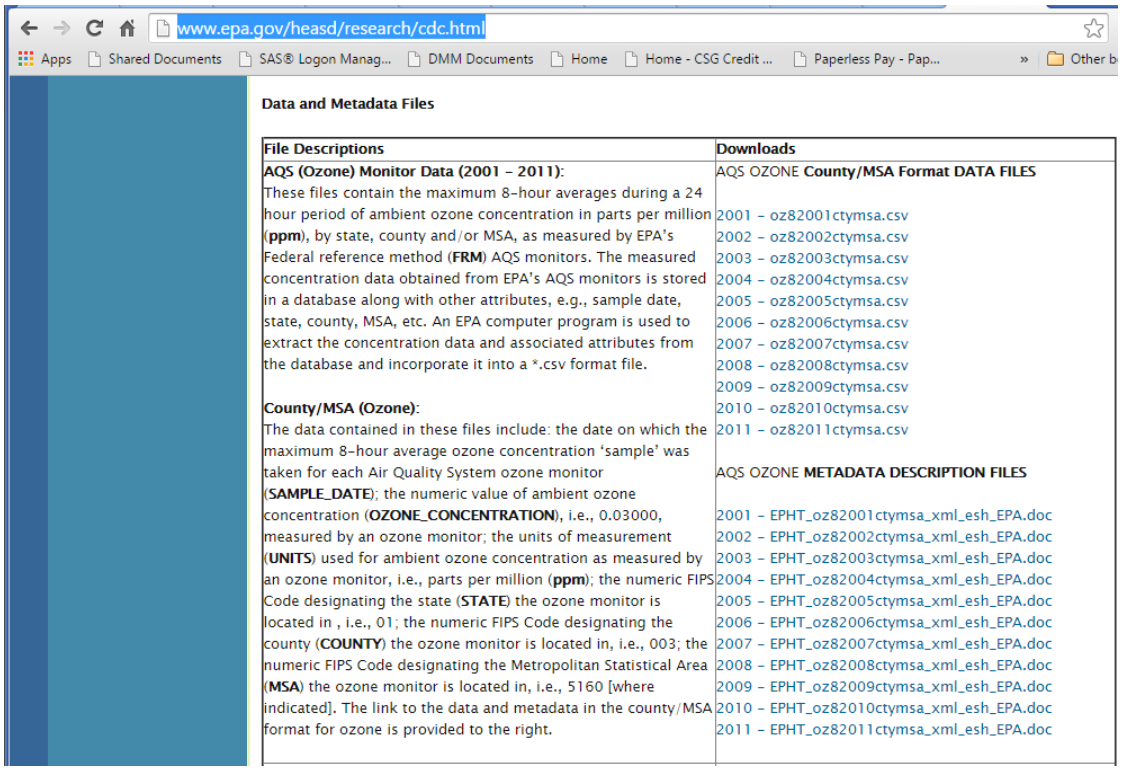


Figure 1. Air Quality System (AQS) Ozone Monitor Data Download webpage

The csv file format data was available by county/MSA was available over the years since 2001 until 2011. A SAS macro program was developed to read the multiple input csv file as shown below and append them together to create one SAS dataset for the time period of 2001 until 2011. Please refer appendix section for the details of the code. Please see below for the sample csv input file.

SAMPLE_DATE	OZONE_CONCENTRATION	UNITS	STATE	COUNTY	MSA
3012001	0.03	ppm	1	3	
3022001	0.029	ppm	1	3	
3032001	0.032	ppm	1	3	
3042001	0.044	ppm	1	3	
3052001	0.052	ppm	1	3	
3062001	0.048	ppm	1	3	
3072001	0.045	ppm	1	3	
3082001	0.065	ppm	1	3	
3092001	0.046	ppm	1	3	

Figure 2. Air Quality System (AQS) Ozone Monitor Data

This file/dataset contains the maximum 8-hour averages during a 24 hour period of ambient ozone concentration in parts per million (ppm), by state, county and/or MSA, as measured by EPA's Federal reference method (FRM) Air Quality System (AQS) monitors. It can be used to provide public health professionals and other users with ambient ozone (air) concentration data for use in assessing the effect of concentration values on health endpoints.

The data contained in this file includes: the date on which the maximum 8-hour average ozone concentration 'sample' was taken for each Air Quality System ozone monitor (SAMPLE_DATE); the numeric value of ambient ozone concentration (OZONE_CONCENTRATION), i.e., 0.03000, measured by an ozone monitor; the units of measurement (UNITS) used for ambient ozone concentration as measured by an ozone monitor, i.e., parts per million (ppm); the numeric FIPS Code designating the state (STATE) the ozone monitor is located in , i.e., 01; the numeric FIPS Code designating the county (COUNTY) the ozone monitor is located in , i.e., 003; the numeric FIPS Code designating the MSA (MSA) the ozone monitor is located in , i.e., 5160 [where indicated].

REFERENCE ZIP TO MSA MAPPING DATA

In order to obtain the actual geography hierarchy information on the AQS data which is at the MSA/FIPS county level, we merge the above AQS data with the standard Zip-MSA (Metropolitan State Area) mapping dataset. The Zip-to-MSA mapping dataset is derived from the SASHELP.ZIPCODE dataset. It is a file containing ZIPCODE level information for the United States including ZIPCODE centroids (x, y coordinates), Area Codes, city names, FIPS codes, and more. The file is indexed on ZIPCODE to facilitate processing, and is updated on a regular basis by SAS. It is provided in transport format so that installations with dissimilar releases of SAS and operating systems can make use of the file. A content listing of the 2005 Q2 file is provided below. SAS has since provided an update to the file as of December 2005, but the contents remain the same. The Zip-to-MSA mapping dataset contains important hierarchal geography information such as the State, City, and Zip code pertaining to different MSA/FIPS in the United States.

zip_msa_2011.sas7bdat

#	Variable	Type	Len
4	areacd	Char	7
26	black	Num	8
6	cbsa	Char	5
8	cbsa_div	Char	5
1	cbsa_title	Char	100
2	city	Char	35
12	city_alias_abbr	Char	8
11	city_alias_name	Char	8
13	city_type	Char	8
15	county_fips	Char	8
3	county_name	Char	35
23	countyarea	Char	8
17	day_svg	Char	8
20	elev	Num	8
24	hhpzip	Num	8
27	hispanic	Num	8
29	housevalue	Num	8
28	income	Num	8
18	latt	Num	8
19	long	Num	8
7	msa2000	Char	5
9	pmsa	Char	4
21	pphouse	Num	8
10	state	Char	2
14	state_fips	Char	8
16	time_zone	Char	8
25	white	Num	8
5	zip	Char	5
22	zippop	Num	8

Figure 3. Zip-to-MSA mapping geography hierarchy dataset

We create a combined dataset for the AQS data by merging the Zip-to-MSA mapping data for our analysis and reporting using the SAS Visual Analytic Dashboard. This concludes the data preparation section of the paper.

SAS VISUAL ANALYTICS POLLUTION DASHBOARD

In this section we will focus on the steps to create visualizations and dashboards using SAS Visual Analytics. We will elaborate how to use SAS Visual Analytics to explore data, to create calculated data items, aggregated measures, define geography items. Create elaborate visualizations such as chats, bar graphs; to graphically visualize pollutant information contaminating the environment; derive hierarchies will be from date, time items and across geographies to allow rolling up the data.

SOURCE DATA PREPARATION

In this section we will explore how to load source data into SAS Visual Analytics server for further exploration, how to create calculated measures and how to define hierarchies.

Load Source Data to SAS LASR server

In order to be able to create analytical reports, we need to import SAS data sets created during the data preparation step which are available on the SAS Application Server and load them to SAS LASR Analytic Server. LASR server consists of the new predefined server (the Public LASR Analytic Server) and the library (the Visual Analytics Public LASR library) and provides broad access and supports the new automated data loading feature to load the SAS datasets in memory.

The final dataset created after merging the 8 hour average AQS ozone data with the Zip-to-MSA mapping dataset is loaded onto the SAS LASR server as follows:

1. Go to the ‘Prepare Data’ tab on the SAS Visual Analytics home screen.
2. Select the ‘Import Remote Data’ from SAS Visual Data Builder
3. Browse through the directories on the server to select SAS data set to load to SAS LASR Analytic Server.
4. Select and load the SAS dataset.

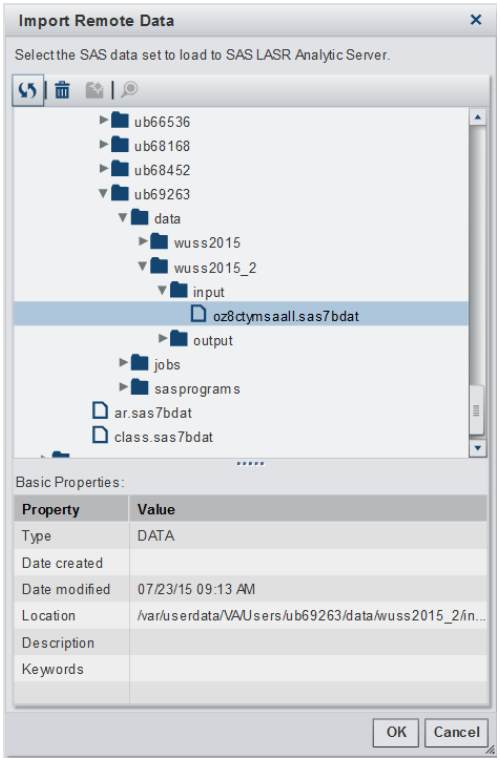


Figure 4. Load source data to SAS LASR server

Create report

SAS Visual Analytics Designer (the designer) enables users to easily create reports or dashboards that can be saved and viewed on either a mobile device or in the viewer. The designer is part of the SAS Visual Analytics product that enables a user with either the SAS Visual Analytics: Analysis role or the SAS Visual Analytics: Administration role can view, interact with, and create reports. Simply open an existing report and interact with the information based on your current needs. Report authors can easily point and click to query central sources of data. Users can drag and drop

tables, graphs, and gauges to create a well-designed report. Users can also add text, images, stored processes, and controls to reports. All of this is accomplished by using the designer, which runs in a web browser. Users do not need to understand a programming language to create reports.

Report authors can easily create reports and dashboards based on data sources that have been provided by a system administrator. They can update reports that were created from visual explorations. Report authors can create reports by importing objects or visual explorations from other reports.

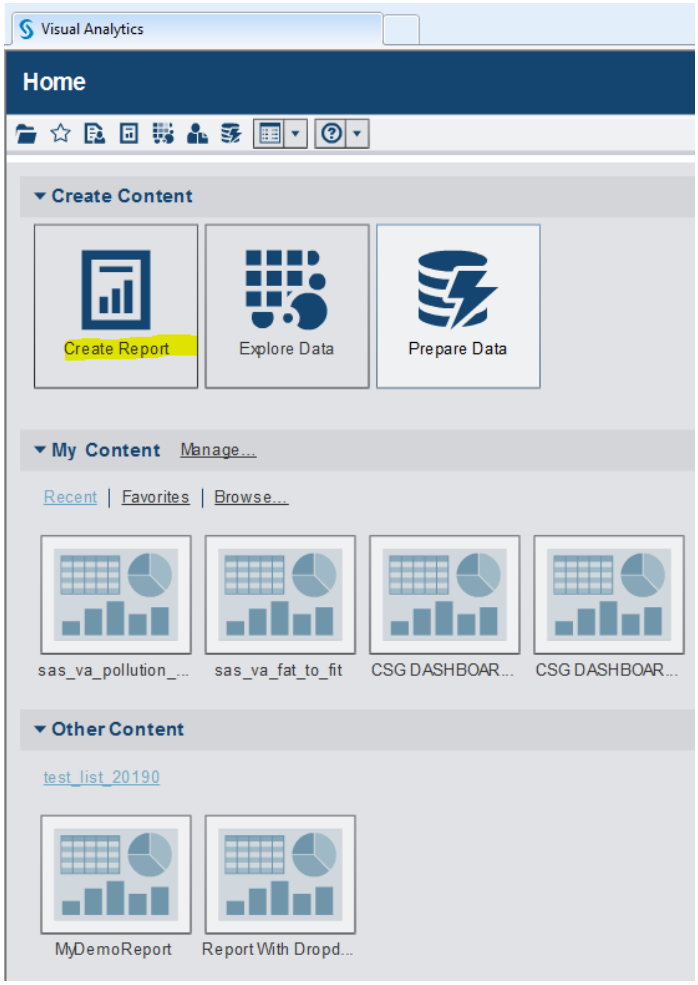


Figure 5. Create report using SAS Visual Analytics

In this section we create the report using the following steps:

1. Go to the SAS Visual Analytics homepage. Select the Create Report tab in the 'Create Content' section
2. Go to Data. Select 'Add data source'. Browse and select the required dataset that you previously loaded in the memory
3. Create any calculated items or hierarchies as required for reporting
4. Add Tables, Graphs, Gauges, Controls or Other texts as desired

Create calculated items

In all the SAS datasets that were imported we had a date value. But in order for creating time series SAS analytics visual dashboards, we were required to create additional measures such as year, month, week from the SAS date available in the source data. SAS Visual Analytics allows us to create these measures by defining calculated items.

The explorer enables you to calculate new data items from your existing data items by using an expression. All calculations are performed on un-aggregated data. The calculation expression is evaluated for each row in the data source before aggregations are performed.

In addition to performing mathematical calculations on numeric values, you can use calculated data items to create

date and time values. For example, if your data contains separate categories for month, day, and year, then you can calculate a date value from each category.

Create hierarchy

A hierarchy is an arrangement of category columns that is based on parent-child relationships. The levels of a hierarchy are arranged with more general information at the top and more specific information at the bottom.

Creating hierarchies enables us to add drill-down functionality to our visualizations. In our process we create a new hierarchy for the State -> City -> Zipcode in order to allow drill-up and drill-down functionality in the report for different geography levels. For more details, please refer the SAS® Visual Analytics 6.2: User's Guide. Please see screenshot below for the hierarchy information.

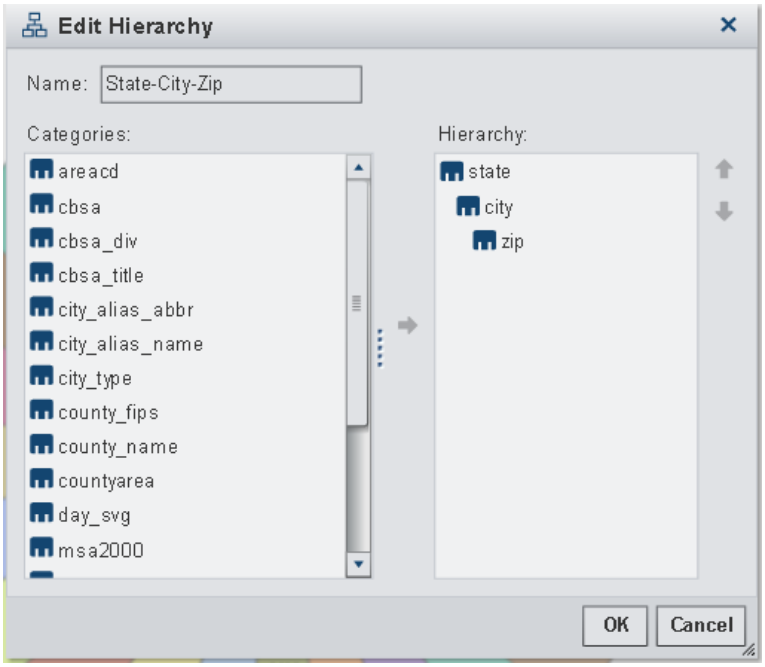


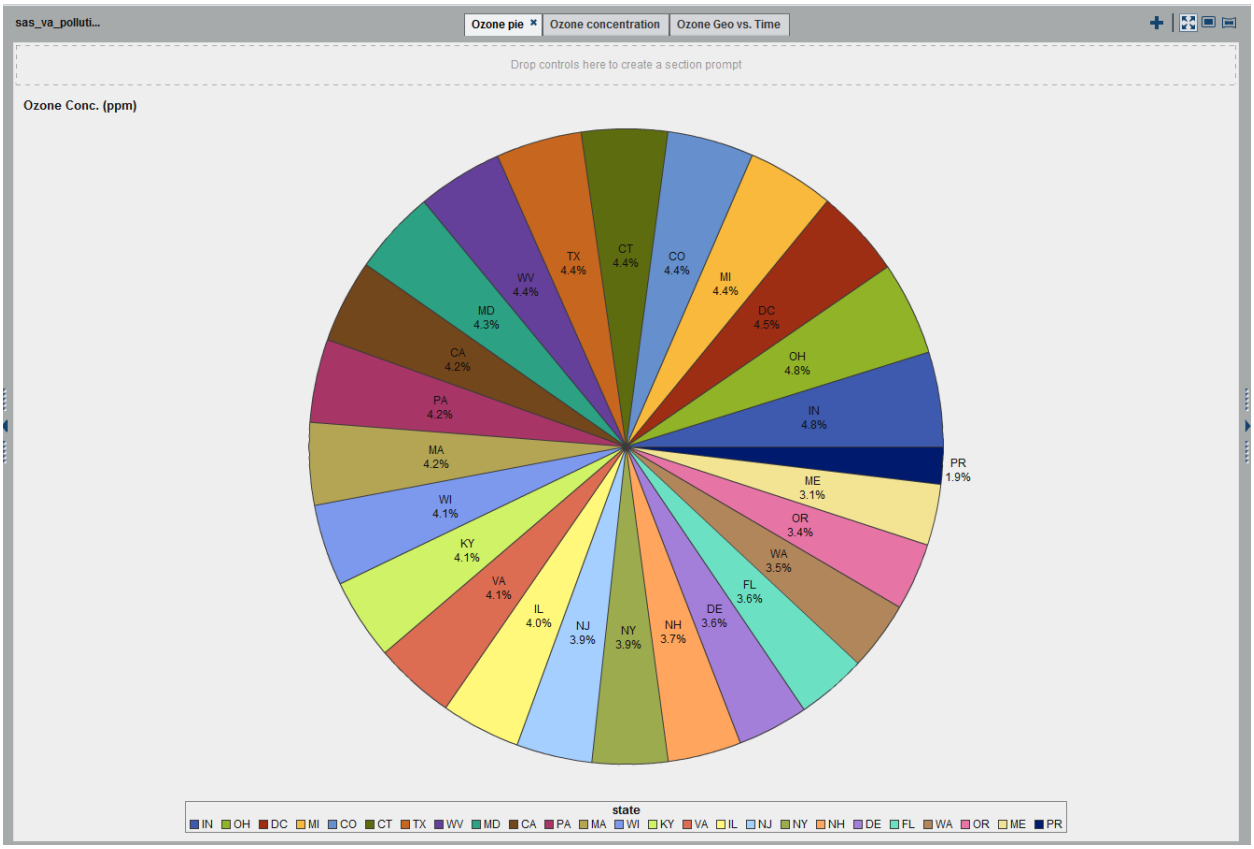
Figure 6. Create hierarchy

DESIGNING REPORTS USING SAS® VISUAL ANALYTICS

In this section we will research creating data visualizations, bar charts, pie charts and other reporting metrics on the sample AQS pollution data using SAS Visual Analytics.

Pie chart

In this section we will discuss how to create a pie chart object in SAS Visual Analytics. Pie chart is a circular chart that is divided into slices by radial lines. Each slice represents the relative contribution of each part to the whole. Pie chart displays a part-to-whole relationship in a circle divided into multiple slices for each value of a category data item based on a single measure data item. Each slice represents the relative contribution of each part to the whole. In a pie chart, the legend is sorted by contribution.

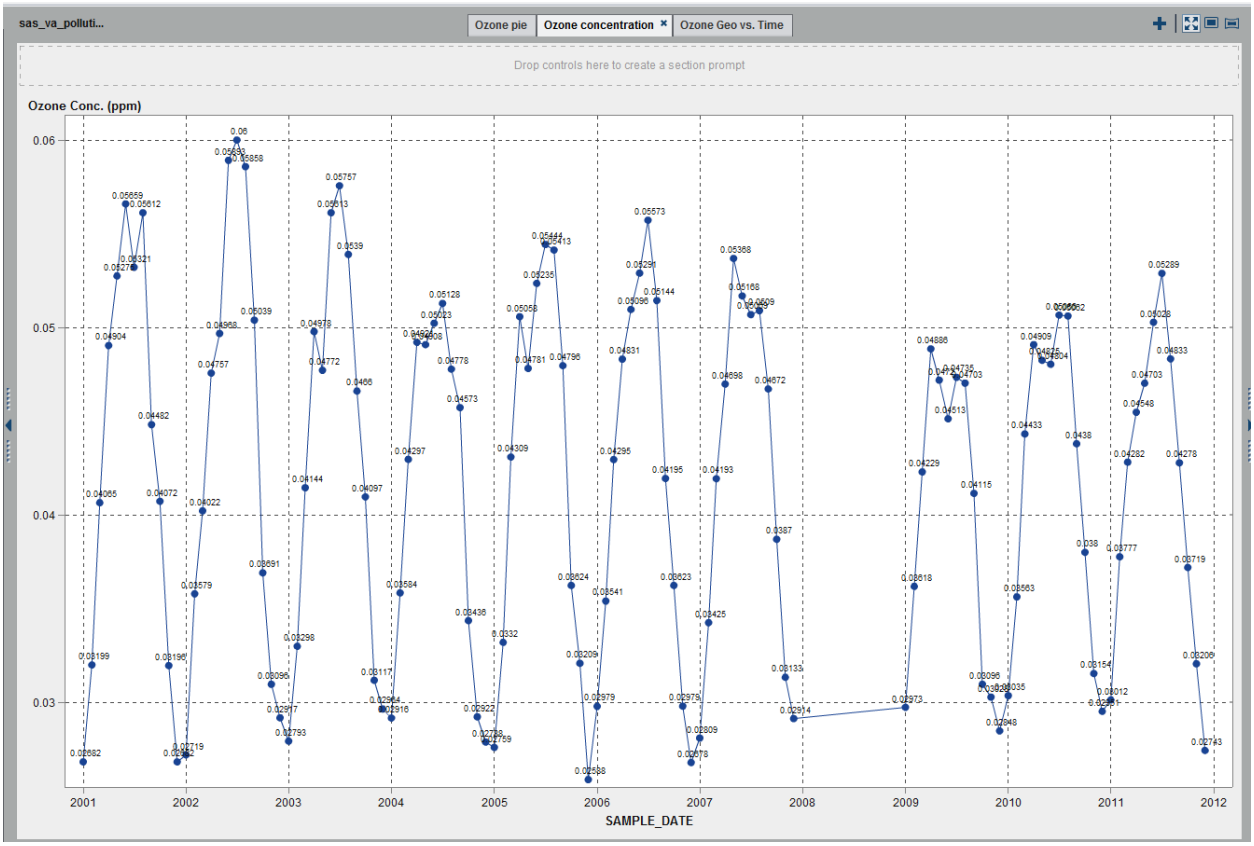


Display 1. Pie chart report

In this example, we create a pie chart by using the AQI data ozone concentration in PPM across the geography hierarchy. Select the OZ8CTYMSA.sas7bdat dataset as dataset object. Drag and drop a pie chart from the objects tab. Select category as State->City->Zip hierarchy and the ozone concentration as the measure. The pie chart object then created allows a drill-up and drill-down functionality on the geography. The data styling can be changed as per user preference to allow different styles for the pie chart object.

Time series plot

A time series plot shows an ordered sequence of values that are observed at equally spaced time intervals. A time series plot requires a date, datetime, or time data item that is continuous. This report monitors the average 8 hour ozone concentration every day over time. The date is plotted on the time axis and the ozone concentration is selected as the measure line. Checking the show markers box in the properties of the Time Series Plot displays the markers and the marker values in the report.



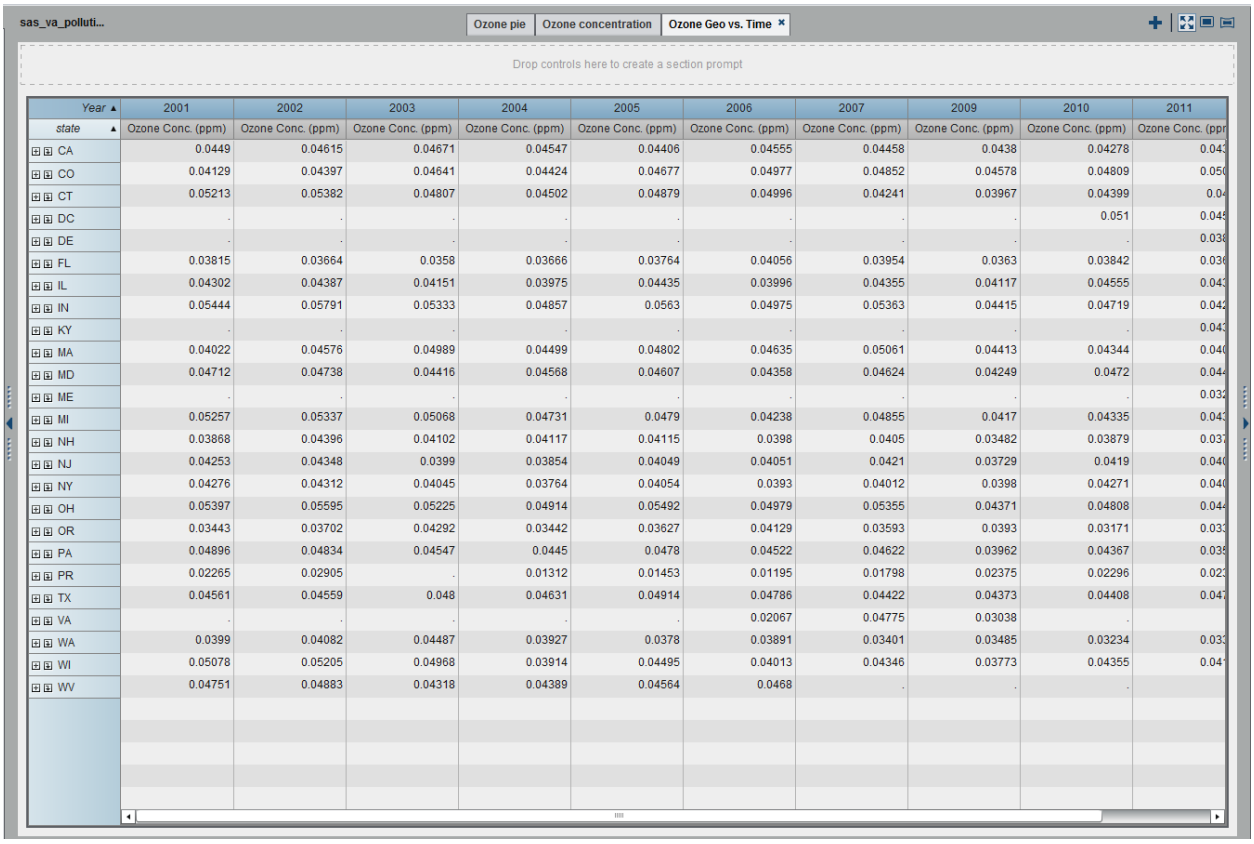
Display 2. Time series plot

Time series plot is created on the source AQI data by selecting the Sample_Date as the time axis and the average 8 hour ozone concentration is selected as the measure. This report shows the ozone: the date on which the maximum 8-hour average ozone concentration 'sample' was taken for each Air Quality System ozone monitor (SAMPLE_DATE); the numeric value of ambient ozone concentration (OZONE_CONCENTRATION), i.e., 0.03000, measured by an ozone monitor in parts per million (ppm). Interesting trend can be observed where we see the ozone concentration goes up and down alternatively in a cyclical pattern.

Crosstab report

In this section we will consider creating a crosstab report to represent the AQI ozone concentration data over time. A crosstab is a two-dimensional table that shows frequency distributions or other aggregate statistics for the intersections of two or more category data items. In a cross tabulation table, categories are displayed on both the columns and rows, and each cell value represents the data result from the intersection of the categories on the specific row and column.

The crosstab report is created by selecting the crosstab object from the objects window. The Years are selected as rows, the hierarchy State->MSA->Zipcode hierarchy is selected as rows and the Ozone concentration in PPM (parts per million) is selected as the measure. The resulting crosstab report allows the drill down capability to view the average ozone concentration across different states in the US, across different Metropolitan State Areas (MSAs) and across different zip codes in the United States.



Year	2001	2002	2003	2004	2005	2006	2007	2009	2010	2011
state	Ozone Conc. (ppm)	Ozone Conc. (ppm)	Ozone Conc. (ppm)	Ozone Conc. (ppm)	Ozone Conc. (ppm)	Ozone Conc. (ppm)	Ozone Conc. (ppm)	Ozone Conc. (ppm)	Ozone Conc. (ppm)	Ozone Conc. (ppr
CA	0.0449	0.04615	0.04671	0.04547	0.04406	0.04555	0.04458	0.0438	0.04278	0.04
CO	0.04129	0.04397	0.04641	0.04424	0.04677	0.04977	0.04852	0.04578	0.04809	0.05
CT	0.05213	0.05382	0.04807	0.04502	0.04879	0.04996	0.04241	0.03967	0.04399	0.0
DC	.	.	.	.	.	.	.	.	0.051	0.04
DE	.	.	.	.	.	.	.	.	.	0.03
FL	0.03815	0.03664	0.0358	0.03666	0.03764	0.04056	0.03954	0.0363	0.03842	0.03
IL	0.04302	0.04387	0.04151	0.03975	0.04435	0.03996	0.04355	0.04117	0.04555	0.04
IN	0.05444	0.05791	0.05333	0.04857	0.0563	0.04975	0.05363	0.04415	0.04719	0.04
KY	.	.	.	.	.	.	.	.	.	0.04
MA	0.04022	0.04576	0.04989	0.04499	0.04802	0.04635	0.05061	0.04413	0.04344	0.04
MD	0.04712	0.04738	0.04416	0.04568	0.04607	0.04358	0.04624	0.04249	0.0472	0.04
ME	.	.	.	.	.	.	.	.	.	0.03
MI	0.05257	0.05337	0.05068	0.04731	0.0479	0.04238	0.04855	0.0417	0.04335	0.04
NH	0.03868	0.04396	0.04102	0.04117	0.04115	0.0398	0.0405	0.03482	0.03879	0.03
NJ	0.04253	0.04348	0.0399	0.03854	0.04049	0.04051	0.0421	0.03729	0.0419	0.04
NY	0.04276	0.04312	0.04045	0.03764	0.04054	0.0393	0.04012	0.0398	0.04271	0.04
OH	0.05397	0.05595	0.05225	0.04914	0.05492	0.04979	0.05355	0.04371	0.04808	0.04
OR	0.03443	0.03702	0.04292	0.03442	0.03627	0.04129	0.03593	0.0393	0.03171	0.03
PA	0.04896	0.04834	0.04547	0.0445	0.0478	0.04522	0.04622	0.03962	0.04367	0.03
PR	0.02265	0.02905	.	0.01312	0.01453	0.01195	0.01798	0.02375	0.02296	0.02
TX	0.04561	0.04559	0.048	0.04631	0.04914	0.04786	0.04422	0.04373	0.04408	0.04
VA	.	.	.	.	.	0.02067	0.04775	0.03038	.	
WA	0.0399	0.04082	0.04487	0.03927	0.0378	0.03891	0.03401	0.03485	0.03234	0.03
WI	0.05078	0.05205	0.04968	0.03914	0.04495	0.04013	0.04346	0.03773	0.04355	0.04
WV	0.04751	0.04883	0.04318	0.04389	0.04564	0.0468	.	.	.	

Display 3. Crosstab report

CONCLUSION

In conclusion, we found that the SAS Visual Analytics is very powerful in-memory data visualization and reporting tool. The data preparation and data transformation can be easily performed using SAS Visual Analytics. New measures required for reporting can be created on the fly. It is perfect for a group of analysts that are not necessarily trained in advanced analytics. It helps the users to visualize results in a quick comprehensive way. SAS Visual Analytics is a great sandbox area to explore, build reports, and share the results with others.

During the course of this case study we accomplished extracting-transforming-loading the third party Air Quality Systems Ozone concentration data in to the SAS Visual Analytics dashboard to create analytical reports on the third party data. We established how easy it is to slice and dice large amounts of data in considerably small amount of time. The drill-down ability by creating hierarchies allows end users to see the data at different granularity across various reports.

It was found that SAS Visual Analytics puts the power of self-service data analysis into the hands of non-statisticians by allowing SAS users to explore the trends within data, create visualizations such as bar charts, time series plots and cross tab reports, and to be able to analyze source data. It was shown how easy it is for the SAS users and non-SAS users alike to create the report using SAS Visual Analytics.

REFERENCES

SAS Institute Inc. 2013. SAS® Visual Analytics 6.2: Getting Started with Exploration and Reporting. Cary, NC: SAS Institute Inc. SAS® Visual Analytics 6.2: Getting Started with Exploration and Reporting. Copyright © 2013, SAS Institute Inc., Cary, NC, USA

SAS Institute Inc. 2013. SAS® Visual Analytics 6.2: User's Guide. Cary, NC: SAS Institute Inc. SAS® Visual Analytics 6.2: User's Guide. Copyright © 2013, SAS Institute Inc., Cary, NC, USA

ACKNOWLEDGMENTS

We would like to profusely thank our organizational management management-John Hagen and Sunmin Park for being very supportive of us to write and present a paper the SAS Global Forum 2016 SAS conference. We would also like to thanks Thomas Billings for his zest, guidance and support offered us to write this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Viraj Kumbhakarna
Enterprise: MUFG Union Bank
Address: 350 California Street
City, State ZIP: San Francisco, CA 94104
Work Phone: 415-273-2413
E-mail: viraj.kumbhakarna@unionbank.com

Name: Baskar Anjappan
Enterprise: MUFG Union Bank
Address: 350 California Street
City, State ZIP: San Francisco, CA 94104
Work Phone: 415 273-2472
E-mail: baskar.anjappan@unionbank.com

APPENDIX

CODE TO IMPORT AIR QUALITY INDEX (AQI) DATA

Following macro code was developed to import the AQR csv files from the EPA website.

```
%macro data_import_ozone_ctymsa(infilename=,dsn=);
DATA &dsn.;
    LENGTH
        sample_date         8
        ozone_concentration   8
        units              $ 3
        state              $ 2
        county             $ 3
        msa                $ 4 ;
    LABEL
        sample_date       = "SAMPLE_DATE"
        ozone_concentration = "OZONE_CONCENTRATION"
        units              = "UNITS"
        state              = "STATE"
        county             = "COUNTY"
        msa                = "MSA" ;
    FORMAT
        sample_date       MMDDYY10.
        ozone_concentration BEST7.
        units              $CHAR3.
        state              $CHAR2.
        county             $CHAR3.
        msa                $CHAR4. ;
    INFORMAT
        sample_date       MMDDYY10.
        ozone_concentration BEST7.
        units              $CHAR3.
        state              $CHAR2.
        county             $CHAR3.
        msa                $CHAR4. ;
    INFILE "&infilename."
           FIRSTOBS=2
        MISSOVER
        DSD ;
    INPUT
        sample_date       : ?? MMDDYY8.
        ozone_concentration : ?? COMMA7.
        units              : $CHAR3.
        state              : $CHAR2.
        county             : $CHAR3.
        msa                : $CHAR4. ;
RUN;
%mend data_import_ozone_ctymsa;
%data_import_ozone_ctymsa(infilename=/var/userdata/PA/COML/Users/ub69263/data/20150722_wuss2015_2/raw/oz82011ctymsa.csv,dsn=raw.oz82011ctymsa);
```

CODE TO MERGE AIR QUALITY INDEX (AQI) DATA AND ZIP TO MSA MAPPING

Following macro code was used to merge the ozone data and zip to msa mapping dataset

```
%macro data_create_ozone_ctymsa;
data input.oz8ctymsa ;
  length state_fips $8. county_fips $8. pmsa $4.;
  set raw.oz82011ctymsa
      raw.oz82010ctymsa
        raw.oz82009ctymsa
        raw.oz82007ctymsa
```

```
            raw.oz82006ctymsa
            raw.oz82005ctymsa
            raw.oz82004ctymsa
            raw.oz82003ctymsa
            raw.oz82002ctymsa
            raw.oz82001ctymsa
            ;
      state_fips=compress(state);
      county_fips=compress(county);
      pmsa=compress(msa);
      drop state county msa;
run;

proc sort data=input.oz8ctymsa; by /*state_fips county_fips*/ pmsa; run;

proc sort data=input.ziptomsa_2011; by /*state_fips county_fips*/ pmsa; run;

data input.oz8ctymsaall;
  merge input.oz8ctymsa (in= a )
        input.ziptomsa_2011 (in= b);
  by pmsa ;
  if a and b;
run;
%mend data_create_ozone_ctymsa;
%data_create_ozone_ctymsa
```