SAS® Visual Analytics dashboard for pollution analysis

Viraj Kumbhakarna
VP – Sr. Analytical Data Consultant
MUFG Union Bank N.A., San Francisco, CA

Baskar Anjappan,
VP – SAS Developer
MUFG Union Bank N.A., San Francisco, CA
Disclaimer

“The contents of the paper herein are solely the author’s thoughts and opinions, which do not represent those of MUFG Union Bank N.A. The bank does not endorse, recommend, or promote any of the computing architectures, platforms, software, programming techniques or styles referenced in this paper.”
**Agenda**

- **Introduction**
- **Data Preparation**
  - Environmental Protection Agency (EPA) Data source
  - Source AQS Ozone Monitor AirData from EPA
  - Reference Zip to MSA Mapping data
- **SAS® Visual Analytics Pollution Dashboard**
  - Source Data Preparation
  - Creating Calculated Items
  - Creating Hierarchy
- **Designing SAS® Visual Analytics Dashboard**
- **Conclusion**
- **Questions**
Introduction

Objective
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

Reports are generated using the open source pollution data available from the US Environmental States Environment Protection Agency (EPA).
Introduction

AirData from EPA.gov

- Data collection agencies report data to EPA via the system called Air Quality System (AQS)
- It makes available several types of aggregate (summary) datasets such as daily/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.
- We attempt to generate time series plots 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 from http://www.epa.gov/airdata/
What Is SAS® Visual Analytics?

SAS® Visual Analytics is an easy-to-use, web-based product that leverages SAS high-performance analytic technologies.

SAS® Visual Analytics empowers organizations to explore huge volumes of data very quickly in order patterns and trends and to identify opportunities for further analysis.
Overview
Overview

**SAS Visual Data Builder**
SAS Visual Data Builder enables users to summarize data, join data, and enhance the predictive power of their data. Users can prepare data for exploration and mining quickly and easily.

**SAS Visual Analytics Explorer**
The highly visual, drag-and-drop data interface of SAS Visual Analytics Explorer (the explorer), combined with the speed of the SAS LASR Analytic Server, accelerate analytic computations and enable organizations to derive value from massive amounts of data…
Overview

SAS Visual Analytics Explorer (contd.)

…This creates an unprecedented ability to solve difficult problems, improve business performance, and mitigate risk rapidly and confidently.

SAS Visual Analytics Designer

SAS Visual Analytics Designer (the designer) enables users to quickly create reports or dashboards, which can be viewed on a mobile device or on the web.
# Introduction

<table>
<thead>
<tr>
<th>Data Input</th>
<th>Data Prep</th>
<th>Design Reports</th>
<th>Dashboards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open source air pollution data obtained from the US Environmental Protection Agency (EPA)</td>
<td>A subset of the AQS data - the Ozone Monitor data from the AQS database was used in case study.</td>
<td>SAS® Visual Analytics designer was utilized to add data source, work with data items, create hierarchies and calculated data items to design reports</td>
<td>SAS® Visual Analytics was utilized to create SAS dashboards for pollution analysis.</td>
</tr>
<tr>
<td>Data collection agencies report data to EPA via the system called Air Quality System (AQS)</td>
<td>AQS ozone data across years is available for download via EPA website</td>
<td></td>
<td>Various visualizations are created using filters, ranking data to create bar/line charts</td>
</tr>
</tbody>
</table>
Data Preparation

- Extract Transform Load (ETL) operations to extract data using SAS® from open source air pollution data was obtained from the US Environmental Protection Agency (EPA) website for creating dashboards using SAS Visual Analytics.

- Explore data feature to gain insight and manage EPA Air Pollution data.

- Load the data on the SAS Visual Analytics LASR server for further analysis and creating detailed reports on the SAS® Visual Analytics dashboard.
Data Preparation

Source AQS Ozone Monitor AirData from EPA

- In this case study we focus on subset of AQS data - the Ozone Monitor data from the AQS database.

- AQS Ozone files contain max. 8-hr. avg. during 24 hour period of ambient ozone concentration in parts per million (ppm), by state, county and/or MSA.

- Ozone data is measured by EPA’s Federal reference method (FRM) AQS monitors.

AQS 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.
Data Preparation

Air Quality System Ozone Monitor Data

• Comma Separated Value (CSV) file format data was available by county/MSA over the years since 2001 until 2011.

• A SAS macro program was developed to read the multiple input csv files as shown append them together to create one SAS dataset for the time period of 2001 until 2011.

• Please refer macro code developed to import data and create data for SAS® Visual Analytics reporting

```sas
%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 Preparation

Zip to MSA Mapping data

- To obtain actual geography hierarchy information on the AQS, we merge AQS data with the standard Zip-MSA (Metropolitan State Area) mapping dataset.
- Zip-to-MSA mapping dataset is derived from the SASHELP.ZIPCODE.
- It is a file containing ZIPCODE level information for the US including ZIPCODE centroids (x, y coordinates), Area Codes, city names, FIPS codes, and more.

### Zip to MSA Merge Code

```sas
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 pmsa;
run;

proc sort data=input.ziptomsa_2011;
    by pmsa;
run;

data input.oz8ctymsaall1;
merge input.oz8ctymsa (in= a )
    input.ziptomsa_2011 (in= b);
    by pmsa ;
    if a and b;
run;
```
Pollution dashboard

- In this section we will focus on steps to create visualizations and dashboards.
- We will elaborate on 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;

Load source data to SAS LASR server

Final dataset created upon merging 8 hour average AQS ozone data with the Zip-to-MSA mapping is loaded on the SAS LASR server as follows:

2. Browse through the directories on the server to select SAS data set to load to SAS LASR Analytic Server. Select and load the SAS dataset.
**SAS® Visual Analytics Pollution Dashboard**

**Create reports**
- SAS Visual Analytics Designer enables users to easily create reports or dashboards that can be saved and viewed on either a mobile device or on viewer.
- 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 admin.
- Report authors can create reports by importing objects or visual explorations from other reports.

2. Create any calculated items or hierarchies as required for reporting. Add Tables, Graphs, Gauges, Controls or Other texts as desired.
SAS® Visual Analytics Pollution Dashboard

Create Calculated Items

- Explorer allows to calculate new data items from existing data items by using 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 e.g., we create calculated items - year, month, week from date value to define a hierarchy.

1. Select SAS dataset used in the report for which you need to create a calculated item. Go to Data, select options and click on ‘New Calculated Item’
2. Enter Name for calculated data item. Select data type from Result type drop-down list.
3. Build the expression for calculated data item by dragging and dropping data items and operators onto the expression in the right pane. For each field in the expression, you can insert a data item, an operator, or a specific value.
SAS® Visual Analytics Pollution Dashboard

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 to add drill-down functionality to our visualizations.
- In our process we create a new hierarchy for: State -> City -> Zipcode in order to allow drill-up and drill-down functionality in the report for different geography levels.

1. Select Data -> New Hierarchy. The New Hierarchy window appears. In the Name field, enter a name for the hierarchy. Select the categories that you want to include in the hierarchy, and then click to add them to the hierarchy.

2. To change order of the categories in the hierarchy, select the category that you want to move. Then, click to move the category up, or down. Click OK to finish creating hierarchy.
Pie chart report

• In this section we will 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 part-to-whole relationship in circle divided into multiple slices for each value of category data item based on a single measure data item.
• Each slice represents relative contribution of each part to the whole.

1. 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.

2. The pie chart object 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 continuous date, datetime, or time data.
- 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.

1. Select 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.

2. The 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.
Designing Reports Using SAS® Visual Analytics

**Cross tab report**

- In this section we create a crosstab report to present 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 crosstab, categories are displayed on both cols and rows, and each cell value represents data result from intersection of the categories on specific row and column.

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

• In conclusion, we found SAS Visual Analytics is a very powerful in-memory data visualization and reporting tool.
• Data preparation and transformation can be easily performed.
• New measures required for reporting can be created on the fly.
• Perfect for a group of analysts not necessarily trained in advanced analytics.
• Helps 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 course of case study we Extract-Transform-Load Ozone concentration data on to SAS Visual Analytics dashboard to create analytical reports on third party data.
• We show how easy it is to slice and dice large amounts of data in considerably small amount of time.
• Explore drill-down ability by creating hierarchies allowing end users to see data at different granularity.
• SAS Visual Analytics puts power of self-service data analysis into the hands of non-statisticians by allowing SAS users to explore trends within data.
• Create visualizations as bar charts, time series plots and cross tab reports to analyze source data.
Questions?
Thank You!
Appendix