**ASSIGNMENT 1**

Using the Sam’s Club DataMart that we created during class on 9/15 utilize your newly found statistical skills to explore and determine the nature of your data in your store. In this assignment you will be asked to create a new project in Teradata Warehouse Miner and then create various analyses that you will store in that project. The professor will look at those within your project himself so there is nothing to hand in for those. You will also be asked to describe your finding as well. In order to do this, you can create a summary Word document named “Assignment 1 LastName FirstName” where you can summarize your findings. As a convenience you can use this MS Word document: A1SummaryStarterDoc that I started for you that will give you a head start. You will submit this document into Blackboard under the Assignments tab into the Assignment 1 ”bucket”.  
   
Complete the following:  
   
1) (10 pts) Determine the statistical variable types for each of the 37 columns in your DataMart (MYSTORE table) and put these into your summary document. Teradata Warehouse Miner will give you a hint as to what it thinks a column is if you hold your mouse over each one. If it's categorical, is it nominal or ordinal? If it's continuous, is it interval or ratio? There may other data types here... refer to the Stats 101 ppt for all possible values.  
   
2) (10 pts) In your Teradata Warehouse Miner Account, create a new project and call it A1. For each of the following columns add a histogram analysis to your project. It’s name should be in the form of  ”Step 2 Histogram Column Name”  
   ■ Transaction\_Date Input/Analysis/parameter/Widths = 1  
   ■ Item\_Quantity  
   
3) (10 pts) For each of the following columns add a Frequency analysis to your project. It’s name should be in the form of  ”Step3 Frequency Column Name”  
   ■ Tender\_Type – For this one in the summary doc, provide the top three tender types and their counts. Use the metadata document to map the number in the DataMart to the actual payment method.  
   ■ Transaction\_Date – No need to include anything in summary for this one.  
   ■ Item\_Quantity - No need to include anything in summary for this one.  
   ■ Primary\_Desc - No need to include anything in summary for this one.  
   ■ Sub\_category\_desc - No need to include anything in summary for this one.  
   
4) (5 pts) Which day of the week has the most items scanned? What is that total for that day for the week (you may have to break out the calculator on this one)? Include this answer in your summary doc. You can use the query below to determine the number of rows of data that come in on each day of the month. If you want you can then also copy and paste the data into excel in order to determine the busiest days.  
   
5) (5 pts) Which day of the week has the fewest items scanned? What is that total for that day for the week (you may have to break out the calculator on this one)? Include this answer in your summary doc.  
   
6) (10 pts) Every card is issued from a particular Sam’s Club. Create and perform a frequency analysis (Called "Step 6 Frequency Issuing Club Nbr" on the **Issuing\_Club\_Nbr** column to see what percentage of items were scanned by people from your store vs other stores. Nothing to include in the summary doc for this one.  
   
7) (10 pts) Where are most of your club members located? Create and perform a frequency analysis on the **Mem\_Zip** column called "Step 7 Frequency MemZip" to determine your top ten zip codes. In the summary doc tell me the top 3 zip codes, the city/state where these are located (<https://tools.usps.com/go/ZipLookupAction!input.action>) and the number of members for each of the 3 zips. Beware, some of this data has been “scrubbed” (made up) so some of the zips may not actually exist so if you run into those, just provide n/a for the city/state.  
   
8) (15 pts) Create Data Explorer and Statistics analysis blocks called "Step 8 Data Explorer" and "Step 8 Statistics" in order to determine the mean, median (the median is in the statistics analysis under the input/analysis tab by another name... hint what percentile is the median value?) and standard deviations for the following variables and in your summary doc, provide a brief description of what this data tells you. If you find something (trust me, there is something odd in one of these columns) that is throwing off your statistics, what is it and how is it affecting your stats?  
   ■ Total\_Unique\_Item\_Count  
   ■ Total\_Visit\_Amt

**ASSIGNMENT 2**

Background Case: AdventureWorks Inc. (AWI) is in the business of manufacturing bicycles and accessories used in biking. AWI has been accumulating data about their customers and all of the sales to these customers in a SQL Server database since July of 2005. For the past few years, whenever a team member needed a report, they submitted a request to their IT dept. and after a few months and numerous meetings they had what they asked for. They have heard all kinds of great things about data visualization and you have shown them your ability to develop ad-hoc reports using Tableau. Your job is initially sift through this data to see if there are any useful “jewels” that can be found in this mountain of data. AWI's hopes are that this information that you present to them will be useful in increasing sales and thus increasing profits in given markets. If this pans out, there will be a big raise and/or promotion in it for you. They want you to create 5 different visualizations and then assemble 4 of them into an interactive Dashboard that he can keep on his desktop for easy/quick access.  
  
They have had their database developers develop the SQL needed to extract out the sales and corresponding demographic data for each sale between 7/2005 and 8/2008 (I realize that this is a couple of years ago but please pretend you are back in 2008). This has resulted in a file containing (filtered) 56,903 records that you will need to conduct your research on.  
   
AWI has 4 specific areas where they think there may be some profitable discoveries and they wants you to explore and document your findings on those scenarios. They also want you to spend a little time looking through the mountain of data for a bit and see if you can unearth any additional jewels on your own.

Complete the following:

* **10 Points -** Work through the steps in the attached word document to bring the data from their data warehouse and then embed it into a Tableau Embedded Workbook (twbx) file by the name of A2 LastName FirstName.
* Take a few minutes to look at the underlying data to see what is captured in your dataset.
* By default, Tableau prings in all numerical data into the workbook as measures. While this is a good idea for numbers that you want to sum, avg, min, max etc. some of the fields are really categorical in nature and ways of categorizing the data. For each of the following fields, drage them from the Measures area of your project up to the Dimensions area.
  + Number of Cars Owned
  + Number of Children at Home
  + Total Children
  + Yearly Income

**Create visualizations for the following 4 scenarios**. For each visualization that you put filters on, make sure to make them into quick filters. **Teacher's Note:** for some of these visualizations, I give you a screenshot of what mine looked like. Your imagination is much better than mine and I therefore give you permission to develop a better visualization than mine. When you do that, make sure to discuss it in your summary document. If you come up with a different visualization and do not note it in the summarization document then we will just have to guess that you were unable to do what we asked :)  
    
**Chart 1 (How did new models affect sales?) (10 points):** Up until the summer of 2006, AWI sold two series of bikes. The 100 series were high-end bikes in either a Mountian or Road configuration and were priced in the $3,500 price range. For the lower-end market they had a Road-650 model that was priced around $700. In Mid-2006, they introduced several new models in the $2,000 range and obsoleted high end 100 series. Develop a line chart that shows how the average Sale amount and the sum of the Sales Amount vary over time. One way to do this would be to have the avg amount on one line and the sum on the other (by ship date). Include and show filters for English Product Category Name and Country Region code. On each row show a different colored line for each English Product Subcategory Name.

**Chart 2 (Global Map of Sales) (10 points)**: If you double click on the Postal Code dimension, you will find out that only your US Postal Codes are correct. Instead Click on Sales Territory Country and City while holding on the control key and then select the symbol maps from the Show Me visualization choice menu. Add filters for Sales Territory Country and English Product Category. Make the sum of the Sales Amount vary the color of the bubbles as well as the size of the bubbles.  
   
  
**Chart 3 (Profitability by Product) (10 points)**: For each row in your dataset you have what the item sold for in the Sales Amount column. You also have the Standard Cost to manufacture that item. You do NOT have the profit which is the Sales Amount - Standard Cost. Create a calculated measure called Profit that gives you this value for each row in your dataset. You also need to know the Gross Profit Margin of each row. For instance, if an item sells for $100 and costs $80 to manufacture the Gross Profit Margin is 20%. It can be expressed as: (Sales Amount - Standard Cost)/Sales Amount. Create a calculated field called Profit Margin that provides this value for each row.

**Chart 4 (Income Level) (10 points):** AWI has purchased demographic data for each of their customers and had the database administrators merge this data into the DW. Your view has several of these fields in them. Two of these fields are Total Children and Yearly Income. They would like to know if all income levels are purchasing their products and if so, does the total number of children in the household make a difference. Before you develop this visualization, think about what you might THINK that it will look like and then compare this to what actually shows up in your story. If there are differences, briefly discuss them in your summary document.

You might try a column chart with these but you may have a better choice. They also want to see how the columns change for each quarter automatically so add the Ship date dimension to the Pages portion and set it for quarters (the second one in the list that will sequentially scroll through each quarter over the entire length of the data.  
    
  
**Chart 5 (you pick) (5 points)**: There are several other Dimensions in the datasource. AWI has given you some direction on the above but this is just their thinking on what may or may not be happening with these dimensions. They would like you to do a little investigating yourself to see if you find any other jewels in the data.  
   
To be totally honest, I am reasonably sure that there **are** some jewels buried in the data but I do not know what they are. Even if you find a combination of groups that do **NOT** reveal anything outstanding that could be just as important. For instance, if you find that Gender makes no difference in the sales for a region then the company can use that information as well. For instance, they would not want to create a gender specific plan for a product if it does not make any difference in the first place :)  
  
Now that I have used Gender as an example you should pick a different combination of columns (dimensions) to slice and dice your data!  
   
**Create a Dashboard (5 points)**: Add a Dashboard to the project and add your chart 5 and 3 of the 4 preceding charts/visualizations that AWI requested. As you drag the charts onto the Dashboard, each chart’s filters will appear on the Dashboard. Also, Make sure the chart name accurately describes the resulting story being “told”.

**Summary Document (15 points):**  
   
Create a 1 page (double-spaced) Word Document "Visualization Summary LastName FirstName.docx" that discusses this project.

This document should

* discuss your intended audience and how the information portrayed in your dashboard helps your audience with their decision-making.
* It should discuss any interesting findings for each of the 5 charts that you created for this project. What is “interesting” you ask? Anything that is inconsistent with the group… i.e. outliers. For instance, if you look at the IncomeGroup chart you will see that there were very few sales to customers in the 40-50K IncomeGroup. Can you think of a reason for inconsistency? Does it represent an opportunity or a threat that the company can react to? Tweak the various filters on the different charts to see if you see combinations of filter values that cause outliers (inconsistencies) to jump out at you.
* If /When you find any interesting results (trust me... there are some interesting things to discuss), make sure to tell us which filter settings caused this information to jump out at you. A screenshot of the visualization is sometimes helpful to do this (but the screenshots do not count towards your one page summary!).

**ASSIGNMENT 3**

Your organization is about to embark on a large marketing effort. To this end, the Marketing director has purchased additional data from an outside source and had this data imported into the organization’s OLTP database. The resulting data is a bit of a mess and the director is unable to get an accurate picture of who lives where and what are their demographic is.  
   
You will utilize one of the most powerful the data cleansing tools on the market today, Microsoft’s Data Quality Services (aka DQS), in order to get this data as clean as possible for proper analysis. If you want to do this on your own Windows machine, DQS is a built-in component of SQL Server 2012 which is available for download through your prepaid DreamSpark account. As a free alternative on any Windows machine this software is available as well via remote desktop into custom-configured servers that I have had made for you (made possible because of our role in the MS Academic Alliance). In case you have forgotten how to configure your remote desktop connection, I have added a tutorial file in the "Remote Connection Tutorials" folder here in the Bb "Weekly Tasks/Course Mat'l" folder.  
  
**Assignment Details:**  
   
Here is the original file containing all of the data that you should start with: "Assignment3.xlsx" (See attachment). Rename the file Assignment3 LastName FirstName.xlsx. In this file, the original data is on a worksheet called A3RawData. Leave the raw data in its original form so that you can go back and look at it if you need to. In this file you should create a new worksheet that is a copy of this raw worksheet and call it A3WorkingData. As you generate DQS output spreadsheets, you should copy the worksheets containing the DQS output (corrected data) into this spreadsheet. You can perform all of the necessary transformations in one DQS data quality project but is OK if you break it up into separate data quality projects for each column if that makes it easier for you. If you end up doing your analysis in several stages that create several DQS output files, that is OK but turn in one file that contains all of the worksheets (raw, working and the various DQS exported output worksheets.  
   
There is bad/inconsistent data in most columns but not all. Use DQS in order to catch what bad data that you can and then copy/paste your clean data use your original data file to make a cleansed data file that you will submit into Blackboard Assignment 3.  
   
Bad data can take a variety of forms and there is some of almost all of the columns in this dataset. Here are some examples of what you may find!

* Inconsistent data – i.e. multiple versions something that means the same thing… Male/M/Man all mean M
* Text where numbers should be – When you set up your domain for fields that should just contain numbers, make sure that you set the domain data type to decimal, or integer (whichever is appropriate). Note that if DQS encounters a text value in column where you have set up the validating domain to be a number, the resulting “corrected” value becomes NULL or empty so you must correct it manually!
* Birthdate must be a realistic date, make sure to set this domain data type to Date and set up rules for the appropriate date range. Note that if a date falls outside of a rule, the resulting “corrected” value becomes NULL or empty so you must correct it manually. You will have access to the original data. For instance, if a birthday is 1/1/1850 and that person’s age is 63, it would be a valid assumption that the birthday should be 1/1/1950 and you should correct the underlying data in your final version
* Empty fields – (Hint: LastName should not be empty) for these you need to go into the domain and look at the domain rules and change DQS\_NULL value from “Correct” to “Invalid”. You can catch these and correct them in Step 3 (Manage and View results page of the DQS Data Quality Project.
* Parsing errors – if you find these, you can either fix them manually or fix them via Excel tools. I will just be looking to verify that they have been fixed.
* Invalid email address – since everyone theoretically has a unique email address, it would not be a good use of time to map every valid/known email in the world into your email address domain! It is a much better idea to just set up a "pattern matching" rule that makes sure that the email address follows a particular pattern of [blah@something.something](mailto:blah@something.something). To do this for this domain, create a domain rule where the

Choose the option of "Value matches regular expression" and input "\b[A-Z0-9.\_%+-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b" (without quotation marks).

All email addresses that do not match this format will appear in the Invalid tab. You should correct them all (I found 11).

* For Zip Codes, make the value a string and you can then check to see if the length of the string is 5 or 9
* Invalid international letters in LastName columns (be careful when cleansing this column – you should not just accept all suggested corrections)
* Invalid Data Type – There are occasionally columns that should be true/false or 1/0 values. In this dataset, those columns have the name “Flag” in them. If you look at the data and there are other data types such as T/F Tru/Fal, or a number other than 0 or 1 then correct it if you can guess what the value should be. For instance, you can assume that T means TRUE Sometimes, you just can’t tell. Thinking back on our study of statistics, what data can you assume for an empty field? You might substitute the average, the median or the mode for this missing data.

You should use the Knowledge Base (KB) that you created in the DQS lab from 3/5 as a starting point for the analysis and subsequent cleansing of your data. You can either create a new KB copied from your lab KB and add new domains to it as needed or just add the new domains to your lab KB. I would probably use the latter of these two choices. When you are adding domains that should be numbers, make sure that you create the domain as the appropriate number type. If there are numbers that are correct and some that are not, make sure that your domains contain these valid values. As I mentioned above, DQS is a very powerful and flexible tool and you may need to add rules to these domains to successfully catch patterns of correct/incorrect data. For instance, it would be impossible to map in all valid email addresses in the world but we can check to see that it follows a certain pattern of name@domain.something.

What if a guy’s last name is “Null”?

Should Total Children be changed to the “reasonable” largest amount?

Should we use the same domains used in the lab? AND will the data from the lab be mixed with the assignment 3 data?

What should non trivial domain values be changed if a value is NULL\_DSQ?

There is Washington, Wash, and WA. Which one represents D.C.? Should it be labeled as D.C. instead?

Is owning 100,000 cars be considered possible? Or should it be adjusted to reasonable largest value?

Should NULL\_DSQ be replaced with something for every domain?? Any global null value I can use? ‘none’?

Data parsing? Where….?

Marital status should just be “Married” or “single”

Make domain boundaries for 2012Sales(-1000000), Age(-200),

BikeBuyer: 2 null

BirthDate: 1 nil

MiddleName: 1 nil(\* 3324)

NumberCarsOwned: 100000, nil (\*2)

NumberChildrenAtHome: 1000 (\* 1), nil (\*1)

PostalCode: (<5) (\* 2) which are 2184, 9227, (>9 ? ) which is 97015-6403

StateProvinceCode: Washington should be WA.

TotalChildren: 100 should be?

US – LastName: nil should be blank