Northwind Data Mining and Statistical Analysis – Data Warehouse

John Wensink

MIS450 - Data Mining

Colorado State University-Global Campus

Dr. Steve Chung

July 5, 2020

Northwind Data Mining and Statistical Analysis – Data Warehouse

The Northwind database is a relational database that uses primary and foreign keys to track and describe a company’s operations across a limited range of categories. Areas such as customers, orders, products, and employees will be the main area of focus for analysis. The data itself is stored in a PostgreSQL data warehouse where it is isolated from the working Northwind database. This ensures consistency over a short period of time as no records in the data will be impacted by the business’ continuing operations and subsequent updates to the production database. Half of the task of data mining is getting the data out of its current location and into our statistical analysis tools. To this end, we can use the Extract Transform and Load (ETL) tool Pentaho by the Hitachi Corporation to extract the data from PostgreSQL, preprocess the data using the provided transformations, and export the data to a file type recognized by our Statistical Analysis Software (SAS) University Edition (SUE). Once the data has been extracted, the other half of data mining is to gain insights into the data. We will be able to use the CRISP-DM model to apply several different data mining techniques including classification, clustering, association, and prediction. Based on this analysis, one will be able to gain meaningful insight into the relationships hidden within the dataset. It can be useful to think of data mining as a process akin to an artist creating a sculpture. First, the slab of data must be imported from its original source, and when it reaches the analyst’s studio layers can begin to be removed in order to bring to light new ways of visualizing the treasures contained within. The same is true with respect to chipping away at the data contained within our database. While data mining is not magic and can not be used to perfectly predict the future, when applied correctly, data mining techniques are able to help one understand the nature of historical data so that he may develop informed plans for the future, and gain a competitive advantage in the field.

**Northwind Data Warehouse**

The first step in completing this assignment was to create a data warehouse for the Northwind database. This was accomplished easily by using the following SQL:

CREATE DATABASE "Northwind\_DW"

WITH

OWNER = Postgres

ENCODING = 'UTF8'

LC\_COLLATE = 'English\_United States.1252'

LC\_CTYPE = 'English\_United States.1252'

TABLESPACE = pg\_default

CONNECTION LIMIT = -1;

The data warehouse was then populated with the Northwind data by entering the SQL directly into the query tool. No issues were noted with the creation of the data warehouse, or the subsequent population of the tables (Appendix A). One of the most important things we can do when exploring a new dataset is to familiarize ourselves with what the data is representing. In this database, it was discovered that the orders tables would likely be used to create a fact table, as it has multiple primary keys including order\_id, customer\_id, and employee\_id. The dimension tables found to be most suitable to include in the star schema include the tables for employees, order\_details, customers, and products. These tables were determined to be suitable for dimension tables as they each have at least one foreign key relating back to the orders fact table (Appendix B).

**Export to SAS**

Now that there is a rudimentary understanding of the data which will be examined, the next step is to get the data from PostgreSQL exported into a format that will be accepted by SAS. To this end, I selected to use Pentaho ETL in order to extract the data from Postgres, transform the data into a format that is compatible with SAS, and finally load the data into a new database in SAS itself. The first step to get this process started is to run the spoon.bat file that directs the Pentaho data integration software to open. A new database connection was initiated with Postgress through Pentaho’s database connection utility (Appendix C). The ETL software has PostgreSQL as a selectable option, and the rest is as simple as filling in the hostname, database name, and username/password. Once the Northwind and Northwind\_DW were shown in bold denoting their shared status, exporting the Extensible Markup Language (XML) file was relatively straightforward as the Pentaho ETL software provides a transformation path that is able to import data from our connected Northwind data warehouse and export the data as an XML file type which is able to be imported into SAS (Hitachi, 2019) (Appendix D). Alternatively, one could save Comma Separated Value (CSV) files for the tables we are going to look at, and use a PROC IMPORT statement in SAS in order to get the data into the program (Stitch, 2020.) (Elam, 2017) (Appendix E). Although Pentaho ETL is a piece of powerful software capable of highly customizable transformations for preprocessing, it really is not needed for this use case, and as such the data was imported using a simple CSV file type into SAS (Appendix F).

**Statistical Analysis**

Now that our data has found its new home inside of a SAS library we are able to start using some of the resources available within the software. To get a broad overview of the data, a tool called summary statistics can be used to get a broad overview of of the data (SAS, 2020) with the ability to create charts and comparative histograms, summary statistics could be used to get a quick snapshot of the age demographics of the employees (Appendix G)

Another tool within SUE that we can put to use is exploratory factor analysis which can be used to aid in classification (Corazziari, 1999). In the example of the Northwind database, we can use a simple distribution model based on price to classify our products into different categories and as such, target customers who are prone to buy more expensive products with targeted advertisements for our high-end products (Appendix H). Some actionable business intelligence that can be gathered from this simple analysis is that although there clearly buyers of the high-cost products (<$260), over 80% of our products are classified in the low-cost category (>$20). Perhaps we should consider offering more products with mid to high costs as they make up a paltry amount of our offerings, and with higher cost items comes higher margins of profits for our organization. I would recommend that we re-evaluate the proportion of low-cost products we offer compared to high and medium cost products. If we were to do an association analysis of the order details tables and run a correlation analysis of the order’s product ID alongside unit price (Appendix J), the output serves to confirm that the vast majority of our business is the sale of low-cost products. An interesting association analysis to run would also be the correlation between an order’s zip code and the total order cost alongside the product ID’s contained within that order.

Clustering is a tool used within SAS that makes use of unsupervised techniques to group together observations that share similar qualities. In the example of the Northwind database, clustering may be of use to understand the relationship between the columns in the orders table. Knowing that we want two or three clusters, we will use the K-means clustering technique to try out all of the columns in the table and check the measures of fit when we select three clusters. The analysis shows that ‘ship-via’, ‘employee ID’, and freight have the lowest r-squared values and as such will be removed from the next iteration. Before removing these two columns and dropping the number of clusters down from three to two, the overall r-squared value was just over 0.5 and does not seem to be a good fit. After making these changes, we were able to improve the fit with a total r-squared value a pleasant 0.759 (Appendix G).

**Conclusion**

This analysis has provided our organization with a clear picture of what we have done in the past, and areas in which we may wish to focus on for the future. Previous operations did not place an emphasis on having someone internal to the company who has a substantial base of knowledge in the required tools inherent to data mining including but not limited to VM Ware, PostgreSQL, Pentaho ETL, LucidChart, and SAS. Moving forward into a company that places a high emphasis on data mining, it is exciting to see that we are already able to make use of knowledge gained this analysis showing that we ought to focus on sales of products that are of a higher overall cost. In the future, I am sure we will continue to apply the data mining techniques we have already explored and continue to build on lessons learned. In the future, It might be beneficial to use a Bayesian Predictor (Snapp, 1996) to make prediction models for the likelihood of a zip code to be higher or lower spenders compared to their peers. This is just a small sample of what our organization is able to do when we are committed to using data mining to form a competitive advantage. It seems there is no limit to what can be accomplished when a data analyst has the creativity and desire to understand his or her data.

References

Corazziari, I. (1999). Dynamic Factor Analysis. Studies in Classification, Data Analysis, and

Knowledge Organization Classification and Data Analysis, 171-178. doi:10.1007/978-3-642-60126-2\_22

Hitachi. (2019, July 19). Use PostgreSQL as Your Repository Database (Archive Installation).

Retrieved June 23, 2020, from <https://help.pentaho.com/Documentation/8.1/Setup/Installation/Archive/PostgreSQL_Repository>

Elam, A. (2017, October 20). Tips for using the IMPORT procedure to read files that

contain delimiters. Retrieved June 23, 2020, from <https://blogs.sas.com/content/sgf/2017/10/20/tips-for-using-the-import-procedure-to-read-files-that-contain-delimiters/>

SAS. (2020). Summary Statistics Task. Retrieved June 14, 2020, from <https://documentation.sas.com/?docsetId=webeditorref&docsetVersion=5.1&docsetTarget=p0aheyslqgh8sxn1lfe9pj883r1m.htm&locale=en>

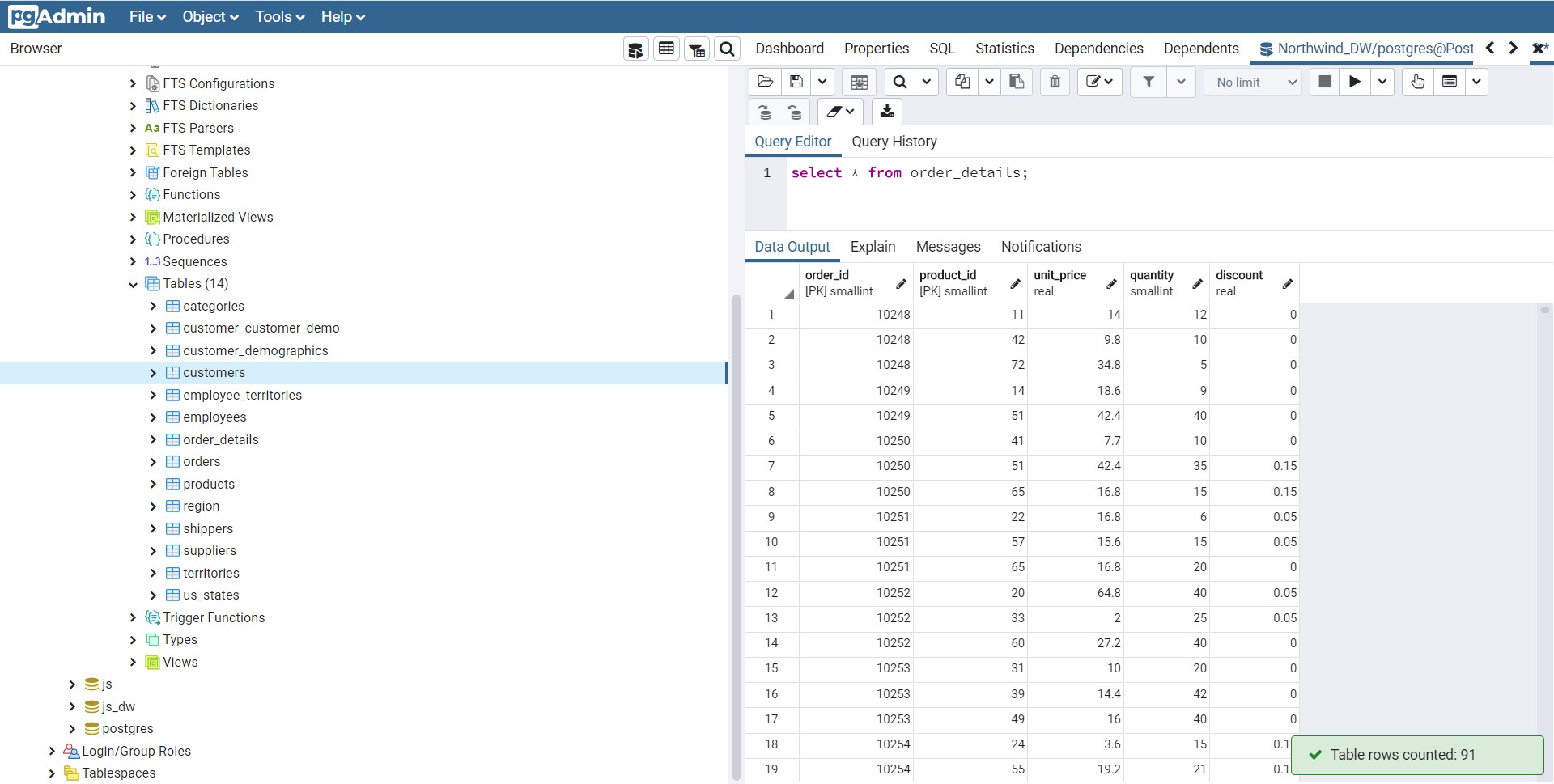
Snapp, R. R. (1996). Predicting the Accuracy of Bayes Classifiers. *Maximum Entropy and*

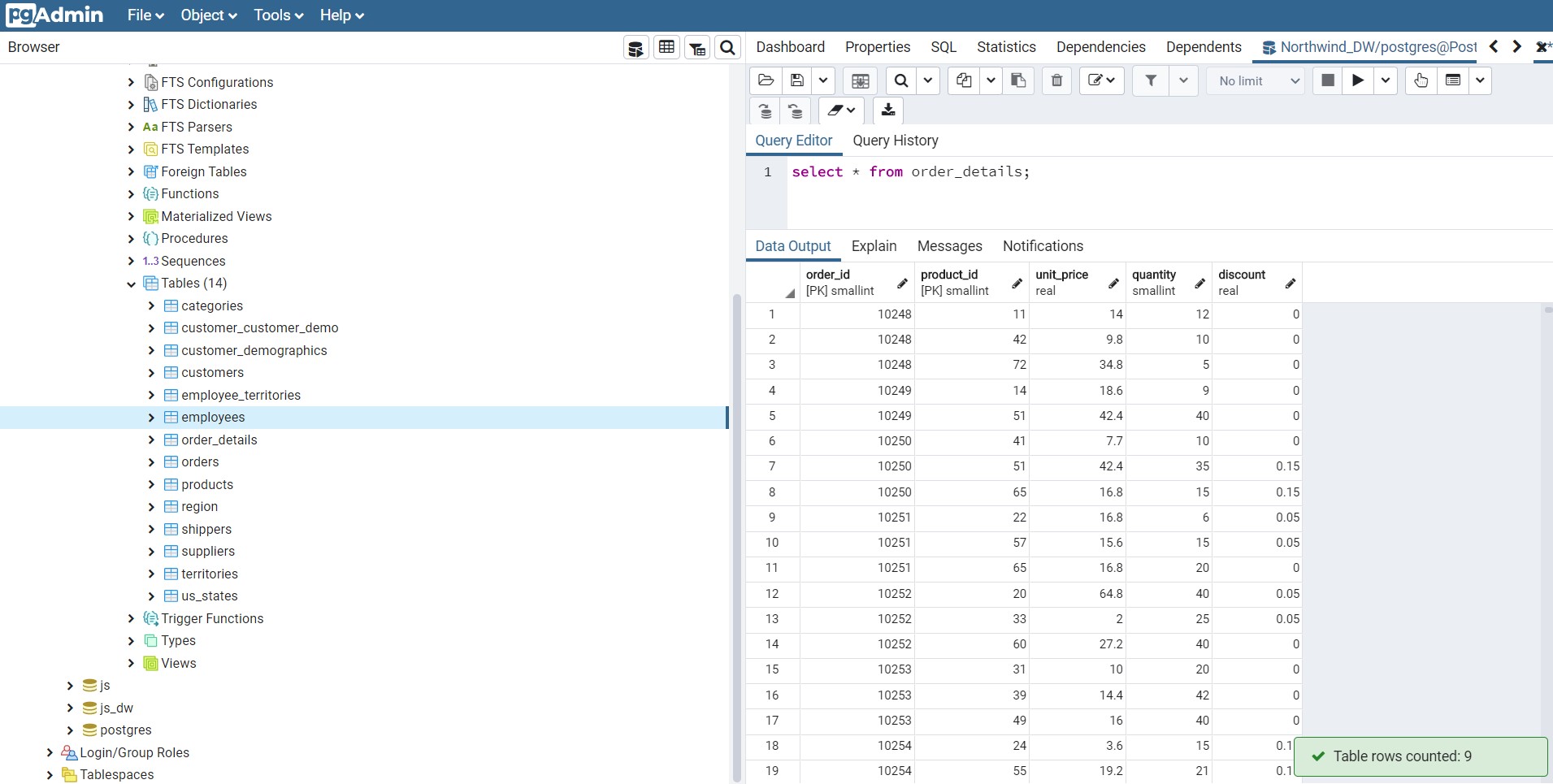
*Bayesian Methods,* 295-302. doi:10.1007/978-94-011-5430-7\_35

Stitch. (2020). ETL your data into your PostgreSQL data warehouse. Retrieved June 23, 2020,

from <https://www.stitchdata.com/data-warehouses/postgresql/>

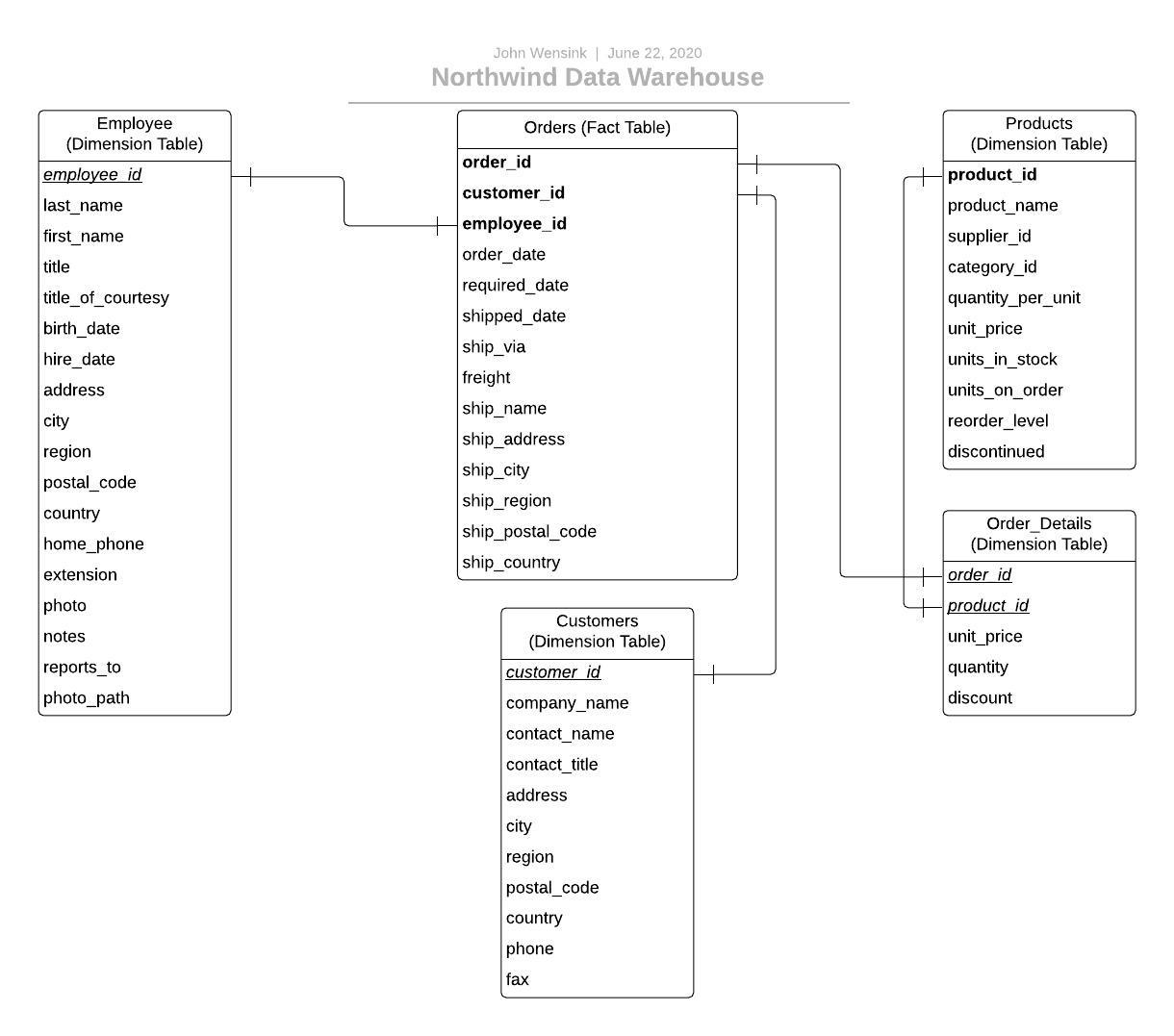
**Appendix A**

*Populated Northwind Tables w/ Row Counts*



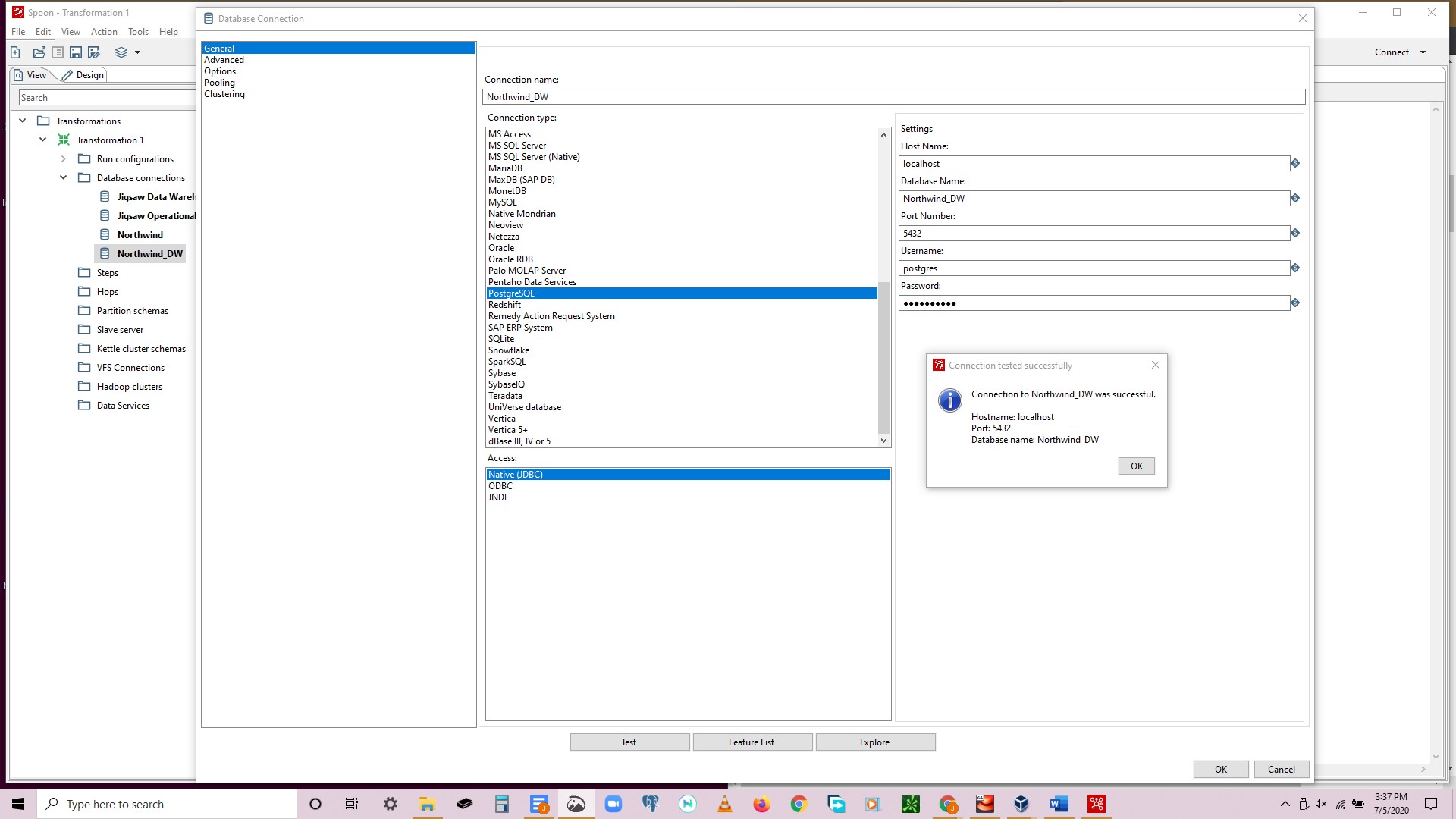
**Appendix B**

*Star Schema w/ Fact and Dimension Tables*



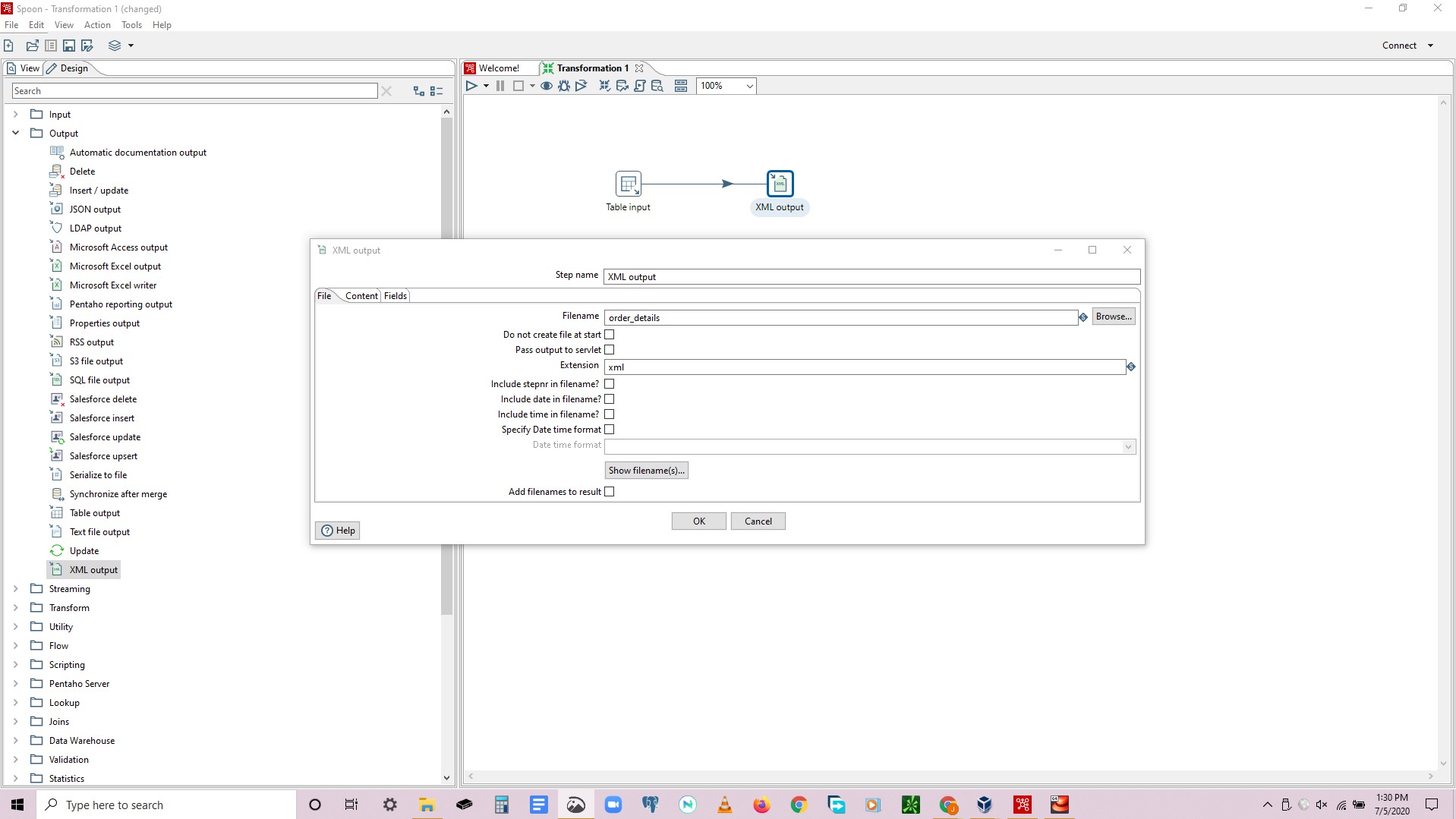
**Appendix C**

*Database Connection with Pentaho ETL*



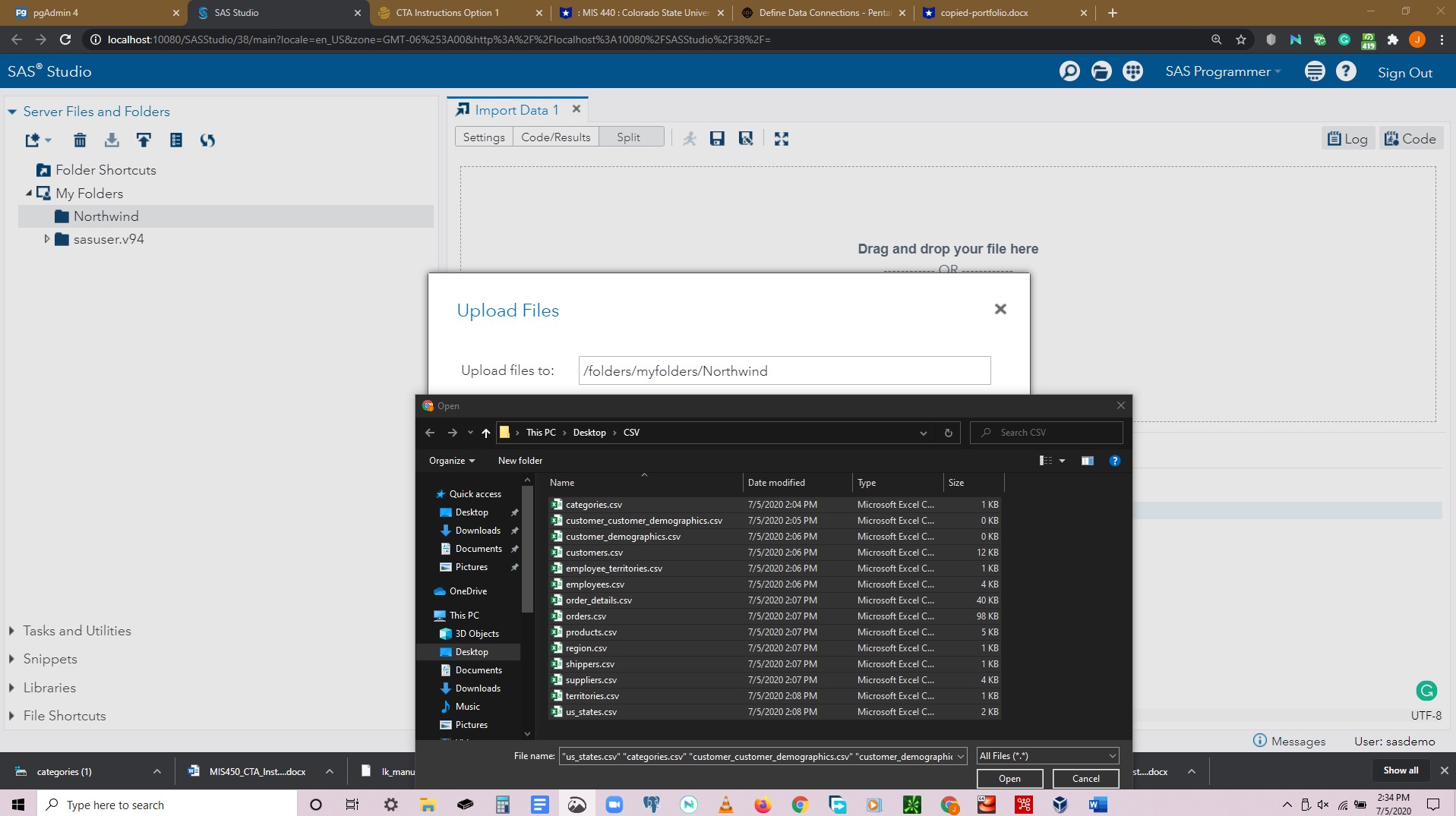
**Appendix D**

*Tables to XML Transformation*



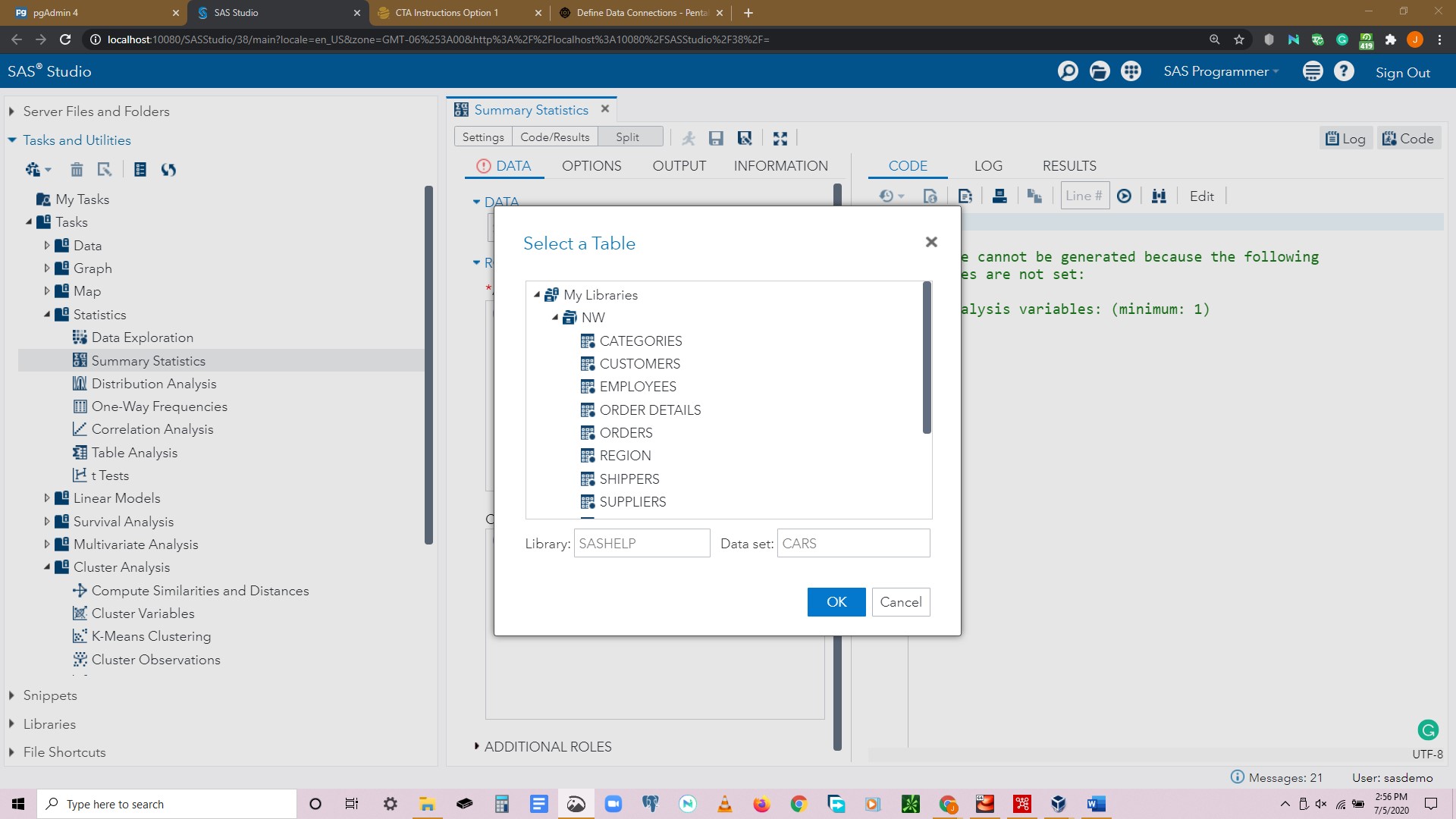
**Appendix E**

*Tables Formatted in CSV*



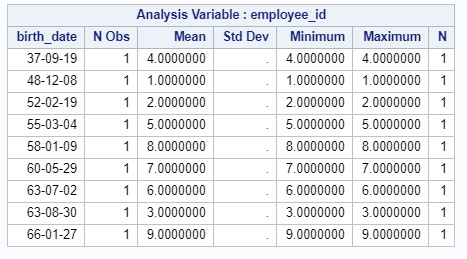
**Appendix F**

*Northwind Database in SAS*



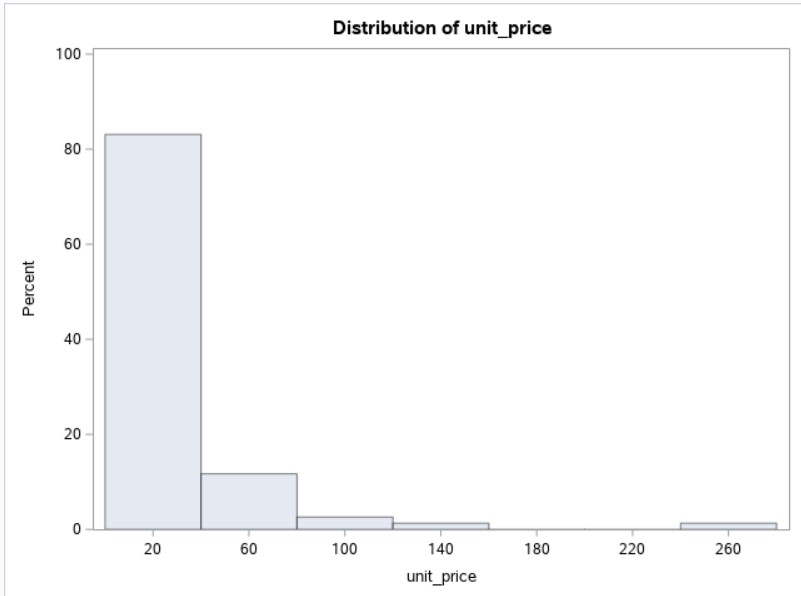
**Appendix G**

*Summary Statistics of Employee ID by Date of Birth*



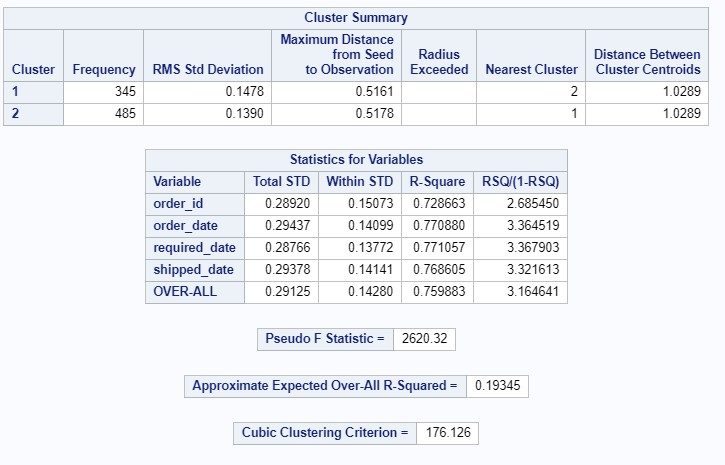
**Appendix H**

*Products Price Distribution*



**Appendix I**

*K-Means Clustering*



**Appendix J**

*Association Analysis*

