**Assignment - Joining Data and Using Dummy Data**

Shaun Pritchard

Rasmussen College

QMB4000

Benjamin Tasker

December 12, 2020

**Assignment - Joining Data and Using Dummy Data**

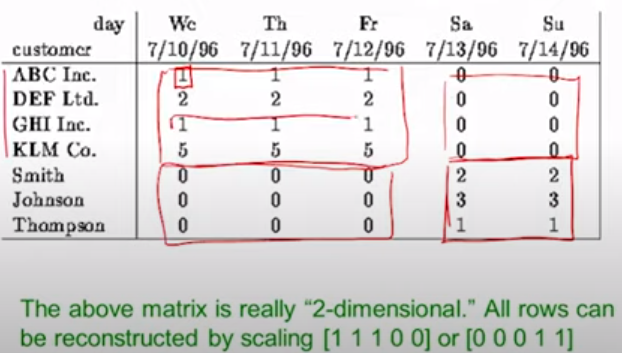
The Data Reduction mechanism can be used to reduce the representation of the large dimensional data. By using data reduction techniques, we can reduce the dimensionality that will improve the manageability and visual ability of data. Allowing us to achieve better more accurate inferences. Essentially making usable smaller sets of data that hold the integrity of the full data set while giving us viable (reduced dimensional) data set to use for analytics purposes.

**Dimensionality reduction can be done in a few different ways:**

* By only keeping the most relevant variables from the original dataset (this technique is called feature selection).
* By finding a smaller set of new variables, each being a combination of the input variables, containing the same information as the input variables (this technique is called dimensionality reduction).
* By removing collinear features.
* By using data reduction techniques such as PCA and ICA.

There are many different data reduction techniques. When we look at a table of data, we can see imagine it as a vector with multiple dimensions based on the columns and rows. Data reduction is used to remove, shorten, or to join the relationships of data to reduce the size and amount of the data set. If we were to look at a dataset with users and basic transactions during the span of a 5 day week as a vector we could see that specific transaction occur at certain times of the week while other users make a transaction at other times. We could reduce the data by breaking down and simplifying our vector to only portraying the 2 different categories of transactions.

We can join and associate columns and rows of data to remove redundant data and associative data. Several of the different data reduction categories are as follows:



* Dimensionality reduction
* Numerosity reduction
* Data compression
* Discrete wavelet transformation
* Principal component analysis
* Nonparametric techniques
* Regression and log-linear
* Clustering
* Sampling

To implement these techniques on a dataset we can implement further techniques on these categories such as:

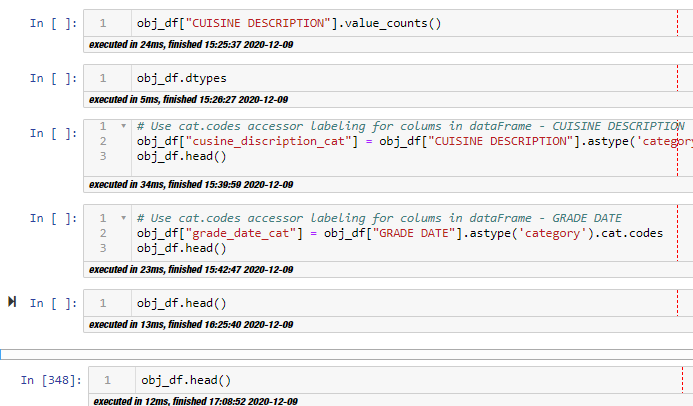
* **Missing Values Ratio** - Data columns with several missing values greater than a given threshold can be removed
* **Low Variance Filter** - Data columns with variance lower than a given threshold are removed
* **High Correlation Filter** - Pairs of columns with a correlation coefficient higher than a threshold are reduced to only one.
* **Random Forests / Ensemble Trees** - Generate a large and carefully constructed set of trees against a target attribute and then use each attribute’s usage statistics to find the most informative subset of features.
* **Principal Component Analysis (PCA)** - Principal Component Analysis (PCA) is a statistical procedure that orthogonally transforms the original n coordinates of a data set into a new set of n coordinates called principal components.
* **Backward Feature Elimination.** - Selecting the maximum tolerable error rate, we define the smallest number of features necessary to reach that classification performance with the selected machine learning algorithm.
* **Forward Feature Construction** - The inverse process to the Backward Feature Elimination.

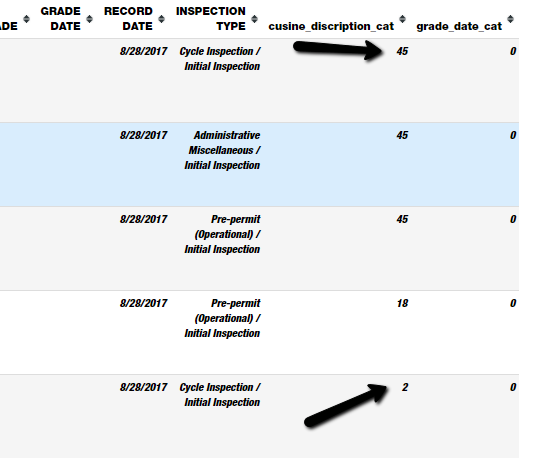
**Explanation of reducing dimensionality ion Airport dataset:**

For our airport data set I simply reduced the data dimensionality by first preprocessing the data set, so we were working with even columns and rows of data. This was accomplished by removing values, using dummy data, and adding average variance to missing and null data. After cleaning the data, it was a matter of combining features with SQL joins and removing features with low variance using SQL queries as follows.

**Reducing The dimensionality:**

After joining the data I exported it to python so I could create a data frame and then create dummy variables (after preprocessing the data) to create a column of data values that would actually represent the categorical data. Using cat.codes accessor labeling in python. I created dummy variables for the **Cuisine Description** and **Grade Dates** columns.





**Can dummy data be incorporated into the dataset:**

Dummy data is meant to impute categorical data. When we have missed, null, or values in error we can use several techniques we must pre-process categorical data by using the most frequent, average recurring features, and dropping null values.

Once imputation is complete, we have several methods we can use to process that data for analysis such as using dummy variables. Using dummy variables means that we can encode the categorical data represented with numerical data like described below so we can get back more accurate results.

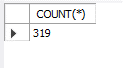
xxx

While using the dummy variable is common to use categorical data and inferences many other different techniques are more reliable and provide more actionable results.

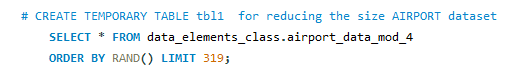
* Listwise deletion
* Imputation of the continuous variable without rounding.
* Logistic Regression imputation
* Discriminant Analysis imputation

Below you will see that we pre-process airport data by using SQL joins and imputing columns such as to use a dummy to represent I call him that is numerical and can be used in inferences. The actual data set is attached in a separate file.

First, I counted the new IATA\_CODE dataset:



I then created a temporary table limited to the size of the IATA\_CODE data set from the main Airport dataset.



I then created a JOIN using the new table created with the IATA\_CODE data using the Airport names columns as the FK to of the IATA\_CODE data set to add the full names of each airport abbreviation together.

By doing this I used another technique of reducing the data set dimensionality.



