Dillard’s is a major retail chain with several stores. Their point-of-sales (POS) data over a period of time is available in <https://nuwildcat-my.sharepoint.com/:u:/g/personal/dkl524_ads_northwestern_edu/Eae3-Uaey_ZNgKKWhwnZ8dwBngaVoXYR1mqd1iN6AEhAlw> The file is over 1GB.

There are 5 tables in the schema. You are encouraged to augment the data with other public datasets.

The data dictionary is provided in a separate file.

Suggested process to undertake:

1. Understand the data
2. Perform data exploration (number of SKUs, number of items per basket, number of stores, most frequently purchased items, busiest stores, etc)
3. Find a machine learning related question to address
4. Feature selection and engineering
5. Modeling
6. Dashboards and story telling
7. ROI – make appropriate assumptions (support numbers used by using the web)

Deliverables

See canvas

Note that the data set is relatively large (not “big data”). If you will have problems querying the entire data set, be creative (in a clever way select a subset of data – based on the knowledge of the problem).