ISSU0053 Data Science and Big Data Analytics

# DATASET INFORMATION

# sales.csv

This dataset contains data from a café, called Executive Express, run by undergraduate business students at a NY public university. It was collected over a ten-week period from January to March 2010.

Variables include a number of types of items sold, along with total sales values.

**SOURCES:** The sales and item data were collected by the café’s student volunteer workers each day from January 19 through March 31, 2010. The daily weather data was obtained from an internet-based historical weather site called Weather Underground, http://www.wunderground.com/history/.

**VARIABLE DESCRIPTIONS:**

A header line containing variable names is included in the data file. Variables are comma delimited.

Each day is one line in the data file. Missing observations are coded as NA.

The variables are as follows:

* Date; starts 19-Jan-10 ends 31-Mar-10
* Weekday: day of the week : Monday, Tuesday Wednesday etc
* Pizza: Number of pizza slices sold per day;
* Wraps: Number of wraps sold per day;
* Pasta: Number of pasta. portions sold per day;
* Desserts: Number of desserts sold per day
* Fruitcup: Number of fruit cups sold per day;
* Chips: Number of bags of chips sold per day
* Juices: Number of assorted bottled juices sold each day;
* Sodas: Number of cups of soda sold each day;
* Coffees: Number of cups of coffee sold each day;
* Total Sales: Total dollar sales indicated on sales register receipt each day;
* Max Daily Temperature (F): Maximum daily temperature recorded for the city each day;
* Min Daily Humidity (%): Minimum humidity level recorded for the city each day;
* Windspeed (mph) Maximum windspeed recorded for the city each day;
* Rainfall (inches) Total daily raindall recorded for the city each day;

SPECIAL NOTES: Days on which the café was closed have NA values for all sales figures

ISSU0053 Data Science and Big Data Analytics

# DATASET INFORMATION

# medical\_data

Located in file: section\_d.Rda

The datasets consist of a study and follow up on a series of patients.

Variables including the number of pregnancies the patient has had, their BMI, insulin level, age, and so on, were collected in a patient study.

The diabetes columns records whether the patient went on to develop diabetes in the following five years.

**Variables in the dataset:**

Pregnancies Number of times pregnant

Glucose Plasma glucose conc. over 2 hrs in an oral glucose tolerance test

BloodPressure Diastolic blood pressure (mm Hg)

Triceps Triceps skin fold thickness (mm)

Insulin 2-Hour serum insulin (mu U/ml)

Mass Body mass index (weight in kg/(height in m)^2)

Pedigree Stores result of calculation related to diabetes among relations. Age Age (years)

Glucagon Measured level of glucagon (ug/ml)

Diabetes Positive / Negative.

Did the subject go on to develop diabetes in the following 5 years.