FN299A: Introduction to Python for Financial Applications

**Midterm**

For your midterm – you will be using a dataset contained in a file called “customer\_orders.csv”. The dataset represents fictitious data containing purchases for particular individuals as well as some demographic information – such as city, state, postal code, address, email address – etc and information about which items were purchased for a particular visit. The items purchased range from Baby Food (column AH ) to Spices (Column BJ). A zero indicates the item was not purchased, a one indicates that the item was purchased.

The file “customer\_birthdates.csv” is a comma delimited text file containing the birth dates for all the customers included in the “customer\_orders.csv” file. Load the dataset and convert all columns to the appropriate data type.

Questions:

1. Describe the purpose of the dataset. Why is the dataset important and how could you use the dataset once the data has been cleaned up (any missing values/invalid data addressed)? (10 points)
2. Describe meaning and type of each attribute in the dataset. What kind of validation could you use for particular columns in the dataset? (10 points)
3. What column would you use to identify individual visits to make purchases in the dataset? How many individual visits are there in the dataset? Show work (use Python). (10 points)
4. What column would you use to identify individuals making purchases in the dataset? How many distinct individuals are part of this dataset? Show work (use Python).(10 points)
5. Verify data quality. Are there any missing values/invalid values? How would you address any missing values (10 points)
6. How many records do you see for purchases inside the US in the dataset? (10 points)
7. How many records do you see for purchases outside the US in the dataset? (10 points)
8. Merge the data frame containing the customer\_orders and customer\_birthdates. Create an excel file with the following columns:

CUSTNAME ADDRESS1 CITY STATE COUNTRY\_CODE POSTAL\_CODE POSTAL\_CODE\_PLUS4 ADDRESS2 EMAIL\_ADDRESS PHONE\_NUMBER CREDITCARD\_TYPE LOCALITY SALESMAN\_ID NATIONALITY NATIONAL\_ID CREDITCARD\_NUMBER DRIVER\_LICENSE CUST\_ID ORDER\_ID ORDER\_DATE ORDER\_TIME FREIGHT\_CHARGES ORDER\_SALESMAN ORDER\_POSTED\_DATE ORDER\_SHIP\_DATE ORDER\_VALUE T\_TYPE PURCHASE\_TOUCHPOINT PURCHASE\_STATUS ORDER\_TYPE GENERATION BIRTHDATE

Make sure that all rows from the customer\_orders dataframe are present. Are there employees without a date of birth? (10 points).

1. From the customer\_birthdates dataset – find out the date of birth for the youngest customer. Also find out the date of birth for the oldest customer. Show work. (10 points)
2. Find out who the youngest individual is. Find out who the oldest individual is. Show work (10 points)

**Bonus Question:** Add a new column called “Age” to the Customer\_BirthDates dataframe, in which you will calculate the customer “Age” based on the birthday. Show work. (Hint: <https://stackoverflow.com/questions/2217488/age-from-birthdate-in-python>) (10 points)

Please create a Jupyter Notebook to submit your work.