Fake Names Canada

Aim:

- The aim of this assignment is just to perform and focus on the ETL process on a fake bank dataset consisting of columns - Gender, GivenName, Surname, StreetAddress, City, ZipCode, CountryFull, Birthday, Balance, InterestRate.

Brief Procedure:

- The given csv file is first converted into a txt file, so that the data can easily be manipulated on Excel. The [Birthday] column is put in a correct date format.The [Balance] column and [InterestRate] column are put in numbers format, the dollar sign and percentage sign are removed from the respective columns. This txt file is then saved as a csv file called ‘FakeNamesCanada.csv’.

- A new project is opened in the Microsoft Visual Studio SSDT. In SQL Server Data Tools (SSDT), an Integration Services project stores and groups the files that are related to the package. For example, a project includes the files that are required to create a specific extract, transfer, and load (ETL) solution. Here, the ‘FakeNamesCanada.csv’ flat file is taken as a source and is converted into OLE DB Destination.

- After running the Data Flow Task, a new database is created ‘RAW_FakeNamesCanada_20190125’. We then build a new working table ‘WRK_FakeNamesCanada’ using Procedure where we use SQL to build and manipulate the working table.

- I was getting errors while building the Working Table due to following reasons:
    1. There were non-numeric values in the column [Balance] and I had to exclude them using the function ‘ISNUMERIC’.
    2. In the [Zipcode] column, the rows with more than 7 characters are excluded using LEN() function.
    3. One of the rows in the [Birthday] column is not in the correct format. It is checked by using ISDATE() function and is excluded as well.

- Additional Exclusions:

Rows are deleted for negative [Balance], rows which do not follow the format of the [Zipcode] i.e. ‘__ __’ are deleted too.