VEHICLE SERVICE

Aim:

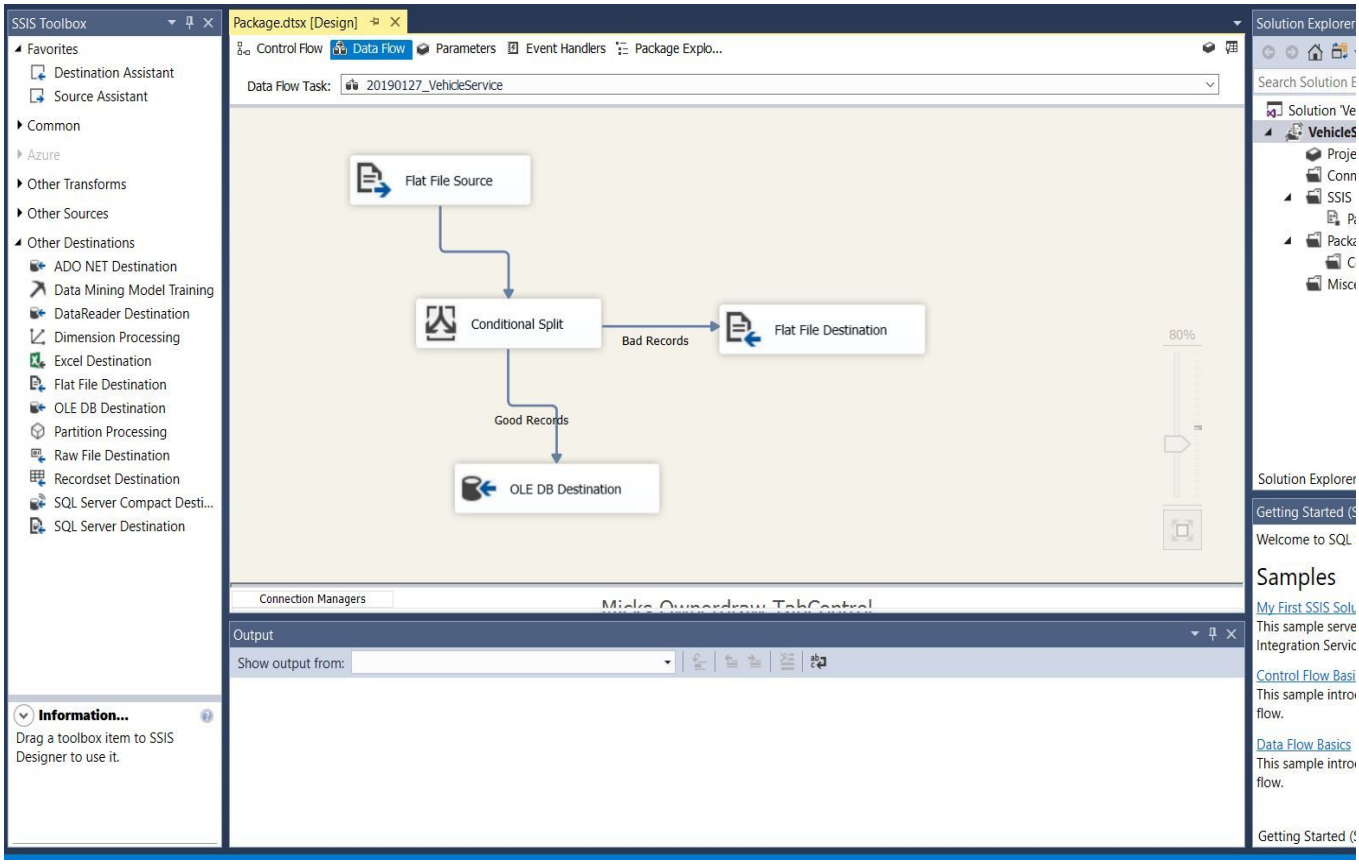
- The aim of this assignment is just to perform and focus on the ETL process on a Vehicle Service dataset which determines the amount of money spent on the car in 3 years. It consists of columns - Customer ID, CustomerSince, Vehicle, 2014, 2015, 2016E.

Brief Procedure:

- The given csv file is first converted into a txt file, so that the data can easily be manipulated on Excel. The [CustomerSince] column is put in a correct date format.The [2014], [2015] and [2016E] columns are put in numbers format. This txt file is then saved as a csv file called 'VehicleService.csv'. However an extra column [Column 6] has been generated because of some error in the csv file.

- 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 'VehicleService.csv' flat file is taken as a source. To eliminate the error which was causing an extra column to occur, we need to put a conditional split with conditions, LEN([Column 6]) > 0 || LEN([2016E]) == 0, if this condition is true for any row, that row will be counted as a Bad Record and will be sent to a different Flat File Destination. Good records are sent to a different OLE DB Destination.

- After running the Data Flow Task, the Good Records are sent to a new Database, 'RAW_VehicleService_20190127'. We then build a new working table 'WRK_VehicleService' using Procedure where we use SQL to build and manipulate the working table.



- I was getting errors while building the Working Table because a row There were non-numeric values in the column [2015] and I had to exclude them using the function ‘ISNUMERIC’.

- I have further worked with anomalies such as duplicates in the [Customer ID] column and finding anomaly in the [CustomerSince] column by checking if there’s any record of older than 100 years. There was one row.