20170410_Batch28_CSE7112c_R Data Preprocessing_Assignment

About data:
Several new airports are opened in major cities, opening the market for new routes (a route refers to a pair of airports) and American airlines has not announced whether it will cover routes to/from these cities. In order to price flights on these routes, a major airline collected information on 638 air routes in the United States. Some factors are known about these new routes: the distance travelled, demographics of the city where the new airport is located and whether this city is a vacation destination. Other factors are yet unknown (e.g., number of passengers who will travel this route). The goal is to predict the airfare on a route (Need not worry about it now. Focus on data cleansing of the given data.).

Perform the following:

- Read the Airfares data in R
- Find the dimensions of the data and report the data types. Check if R is reading the data type as desired. If not then convert it to the right data type.
- The missing value is denoted as *. Identify how many missing values are there in data. If ignored, how many records will be lost. Find the subset of data where you have complete data.
- Clean the data: remove any special characters in the data and check if the data types are consistent.
- What is the average fare if the city is a vacation destination when compared to if the city is not a vacation destination
- The cities columns have both city and state names. Split them into city and state names
- How does the average fare differ if the slot is free or controlled?
- How does the average fare differ if the gate is free or constrained?
- Does the average fare vary by starting city or by destination city?
- Do you find any outliers in the data?
- Bin the starting city and ending city’s income into 10 levels
- Standardize all the numeric data or bin them
- Using visualizations, identify best insights from this data either using the raw data/processed data. Show atleast 2 insights that are of business value