

**EEC 525 / CIS 660 DATA MINING**

**LAB 1**

**Part 1 – Feature Selection & Part 2 – Data Pre- processing and Transformation**

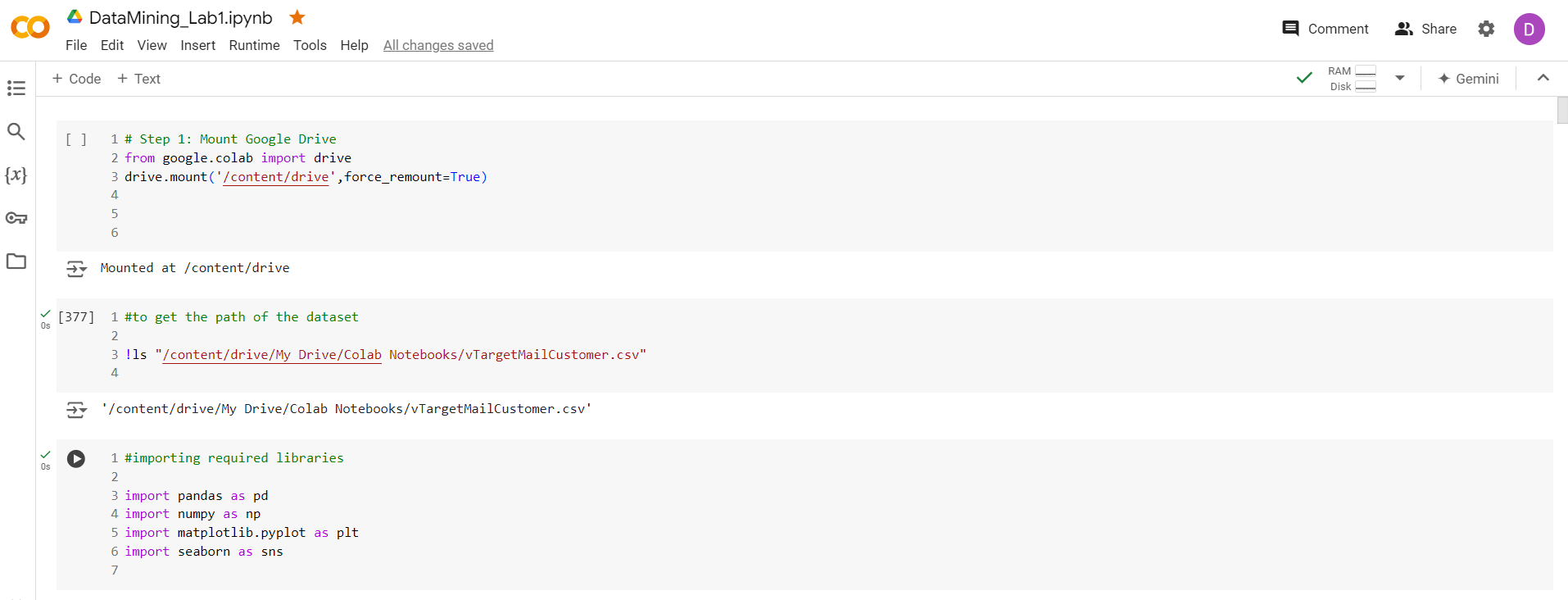
**Submitted By**

**Dinky Mishra**

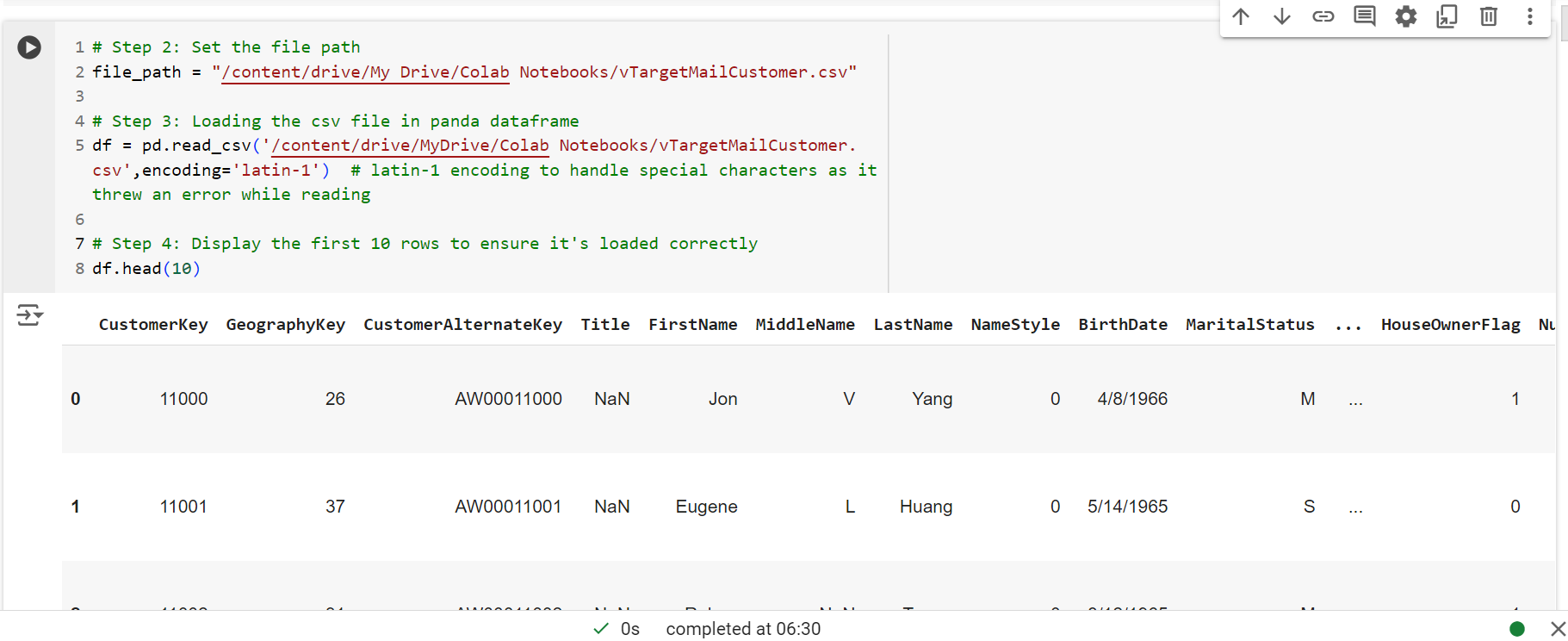
**CSU ID: 2864923**

**Table of contents**

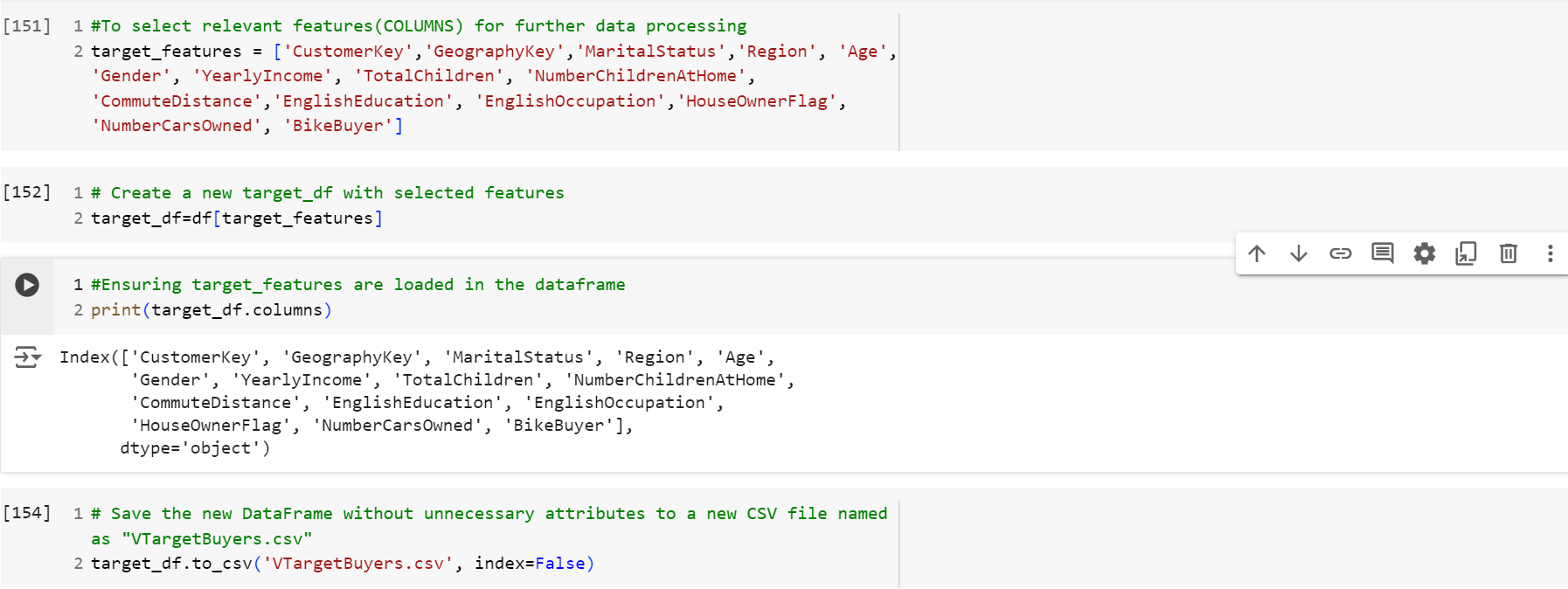
|  |  |  |
| --- | --- | --- |
| **S.NO** | **Content** | **Page No** |
| 1 | PartI: Importing data set | 3 |
| 2 | Feature Selection & saving the file | 5 |
| 3 | Finding properties of selected features | 7 |
| 4 | Part 2 : Handling Null values | 9 |
| 5 | Perform Random Sampling, Bootstrap Sampling with Replacement | 10 |
| 6 | Mean/Variance/Standard Deviation for Ordinal Numeric attributes | 11 |
| 7 | Identify Data Dispersion (Distribution) for each feature using one of the methods: Histogram | 13 |
| 8 | Identify Outliers with 1.5IQR Method to Replace | 18 |
| 9 | Discretization (Binning/Histogram) on Continuous features or Numeric features with too many different values | 20 |
| 10 | Get Q1, Median, Q3 of the Grouped data (those features that were transformed into Histograms by Discretization) | 21 |
| 11 | Identify which feature needs to be transformed by Binarization (One Hot Encoding) to apply Binarization for those features | 22 |
| 12 | Normalization or Standardization | 23 |
| 13 | Saving the final processed data in a new csv file | 24 |

Part 1: Feature Selection: 1. Examine each attribute in the data file (vTargetMailCustomer.csv ) to select a set of Features (Attributes) that would affect to predict future bike buyers. Remove all the unnecessary attributes from the data file VTargetBuyerMailList to Create a new file VTargetBuyers with the selected Features only. Include ID and the Class column (BikeBuyer) in your file. However, the class attribute BikeBuyer should NOT be included for the rest of the data processing and the calculation below.   
  
For the lab,I have used Google Colab platform.   
Steps to load the csv format dataset to a dataframe:   
STEP1: Mount Google Drive to make sure the required dataset is stored in the google drive while loading. Checking the path for required dataset  
Importing the required libraries for the lab  


STEP2: Set the file path  
STEP3: Loading the csv file in panda data frame  
STEP4: Ensuring that the dataset is loaded correctly in the dataframe

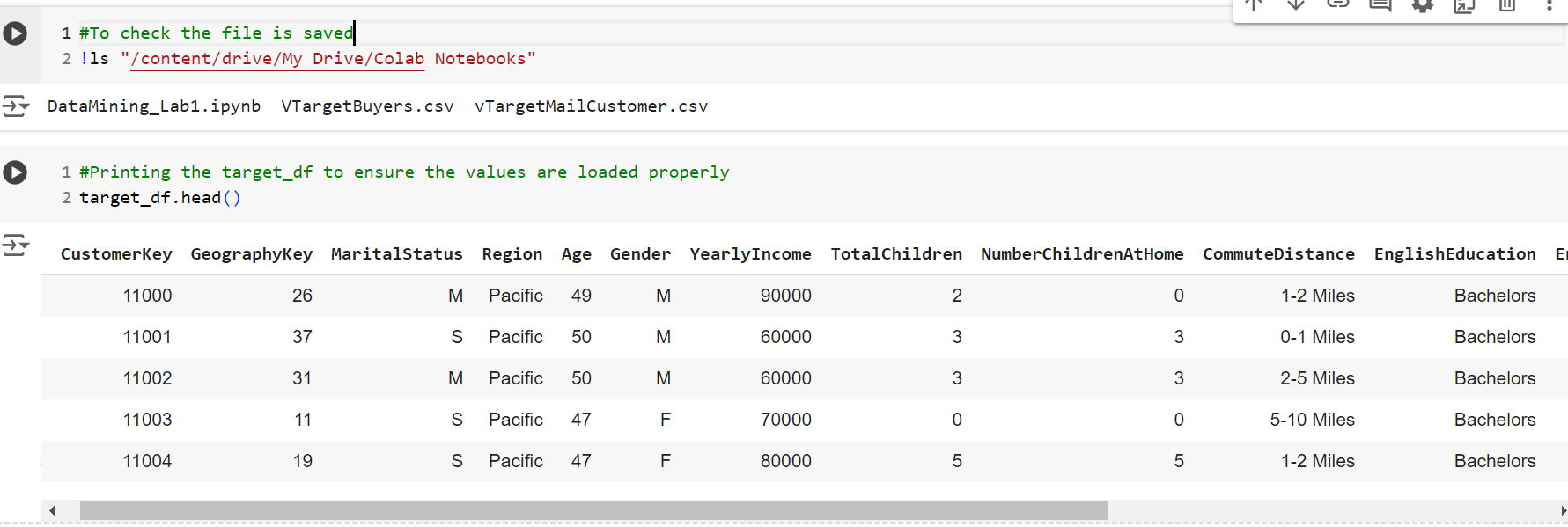


Feature Selection & saving the file:

Now ,  
Select the target features and create a new file named as “VTargetBuyers.csv” and save it.

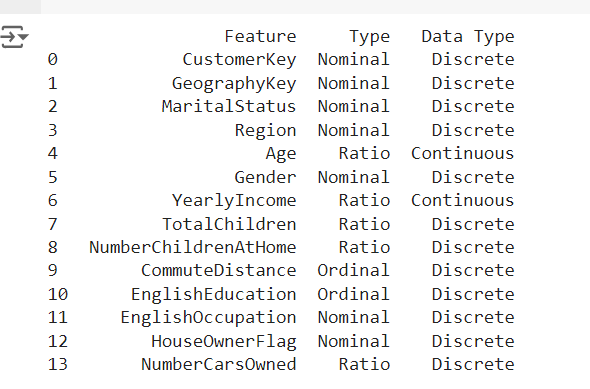


To check the file location of the dataset and to print the target\_df to ensure that data is loaded properly.



2. For each selected feature in your selection, determine the Properties of data covered in class as follow: 1) All the Properties Whether it is Nominal, Ordinal, Interval, or Ratio 2) Data Type whether it is Discrete, or Continuous

For part 1 & part2 of the lab,we need to perform the tasks by excluding the column “BikeBuyer”   
Therefore target features only include the below columns:

Now for selected features, properties are determined below :  
OUTPUT:  


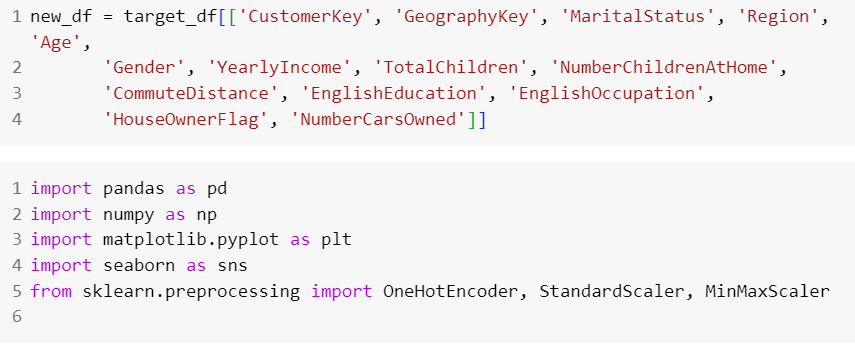
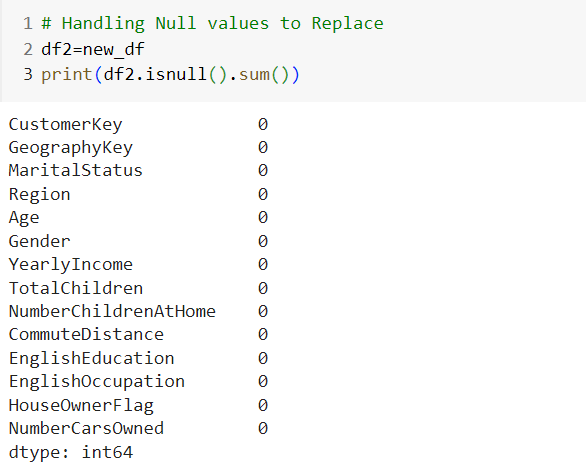
Part 2: Data Preprocessing and Transformation

Use all the data rows (~= 18000 rows) with the selected features as an input file to apply the tasks below, do not perform each task on the smaller data set that you got from your random sampling.

For each selected feature in Part1,

* Handling Null values to Replace

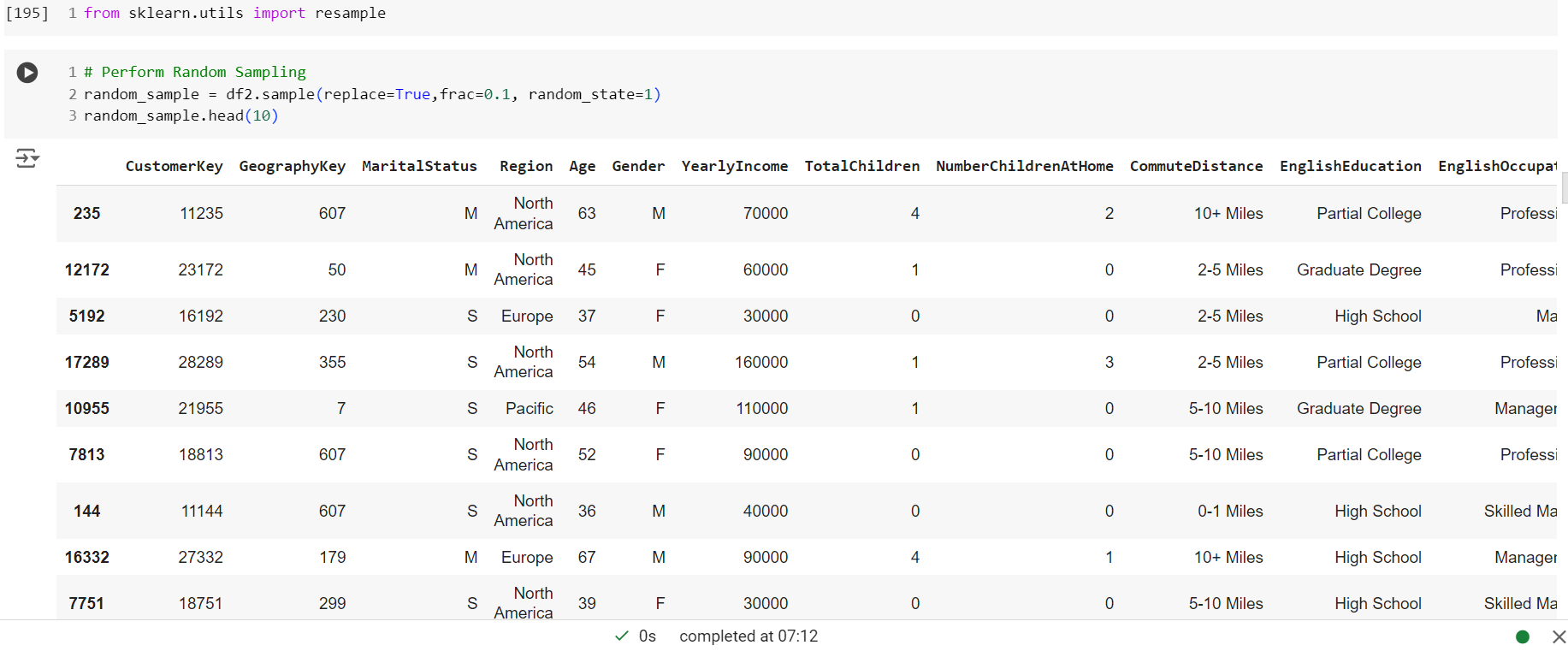
The new dataset is loaded in dataframe new\_df excluding the BikeBuyer column and necessary libraries are imported

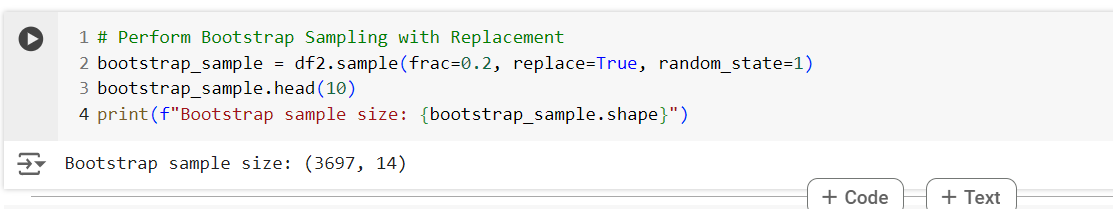
For handling null values, as we could see there are no null values so there is no need to replace it .  


* Perform Random Sampling, Bootstrap Sampling with Replacement

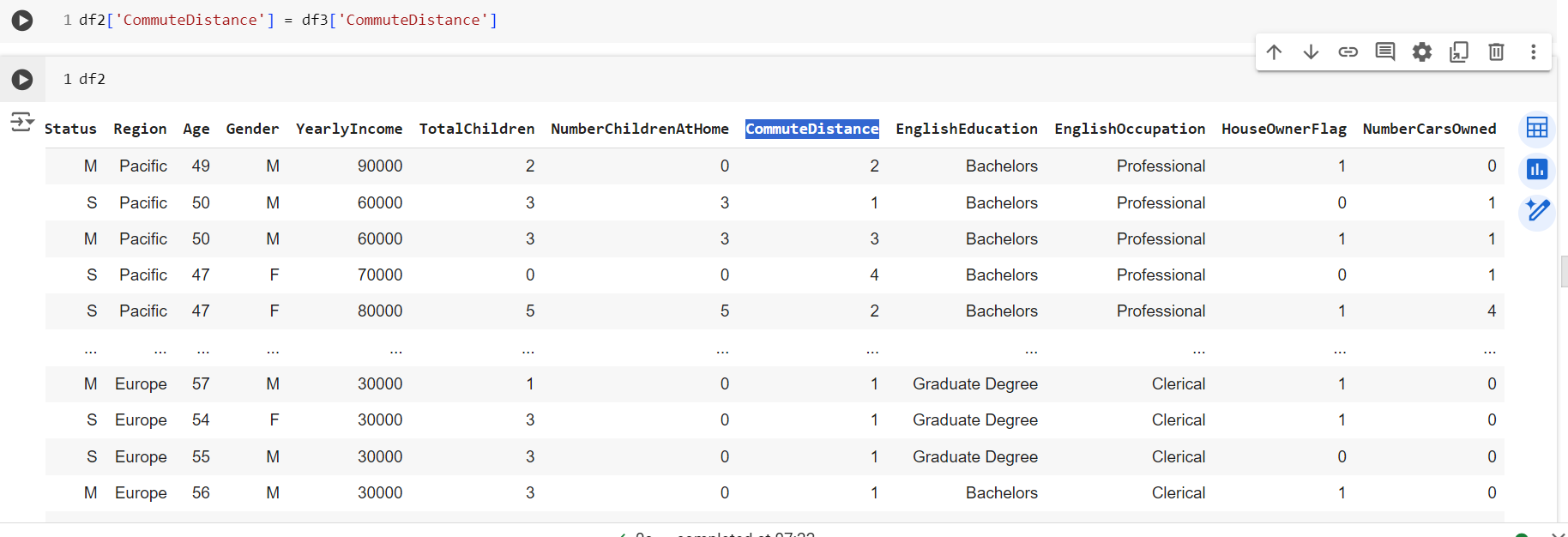
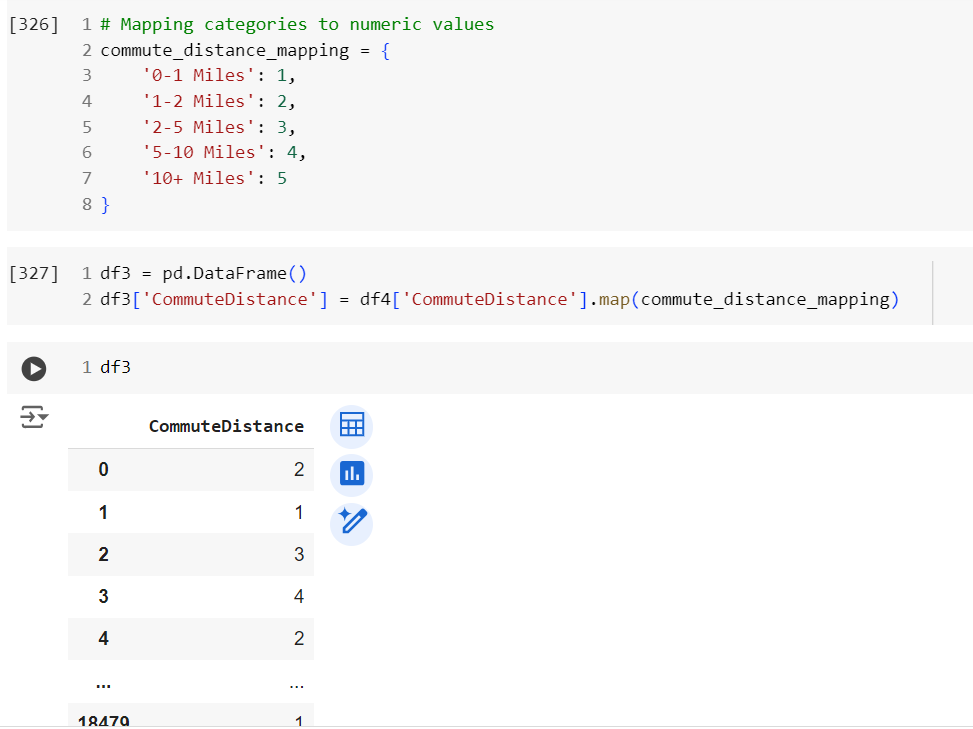
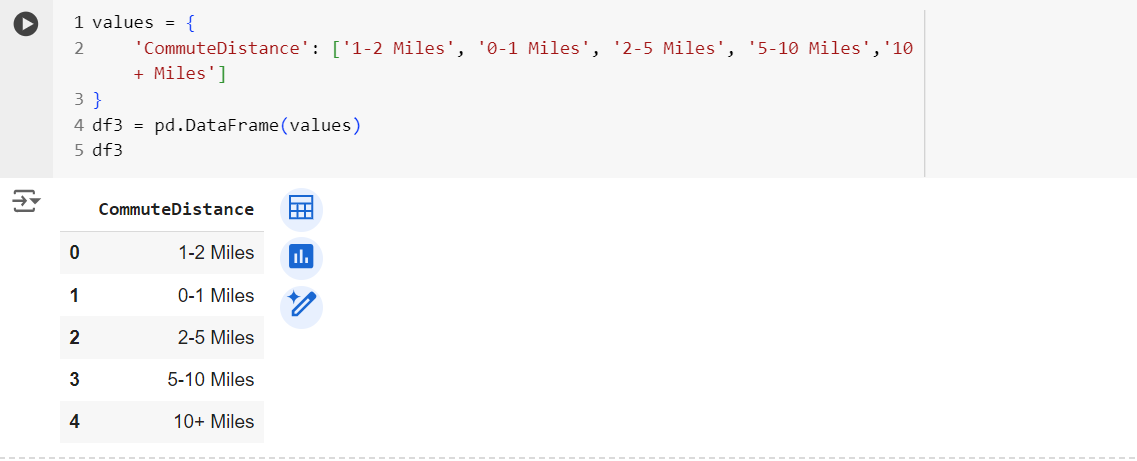
For Sampling we need the necessary resample libraries which is imported below.

Random Sampling 10% with replacement

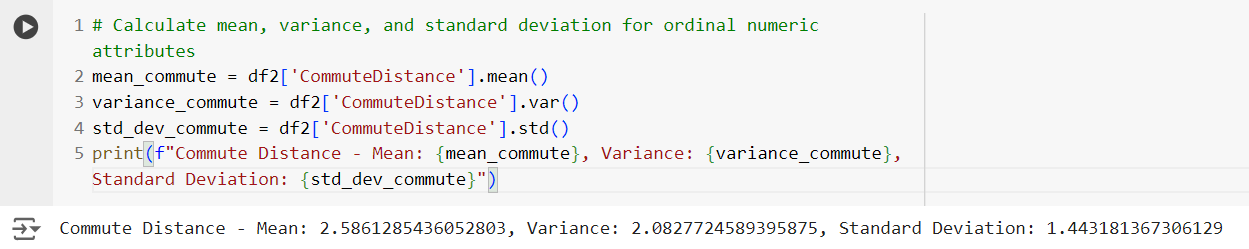


Bootstrap Sampling with 20% replacement  


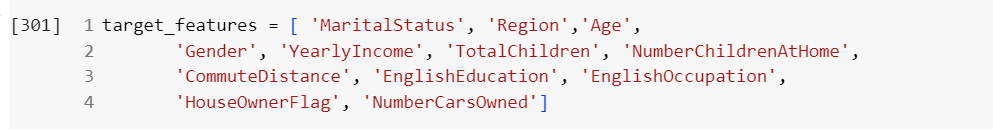
* Mean/Variance/Standard Deviation for Ordinal Numeric attributes

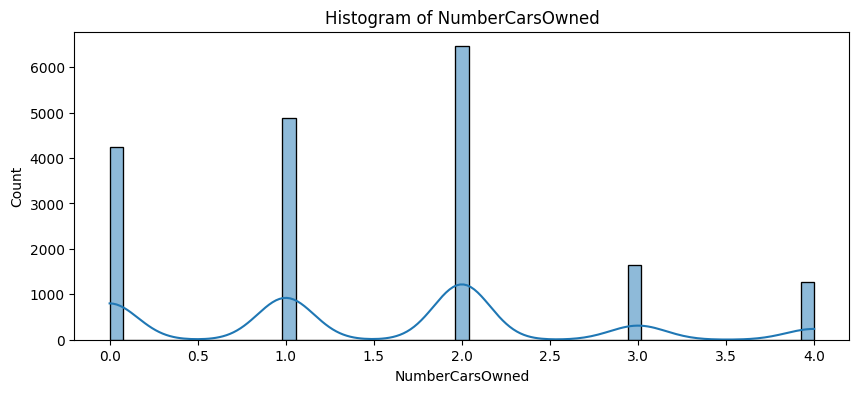
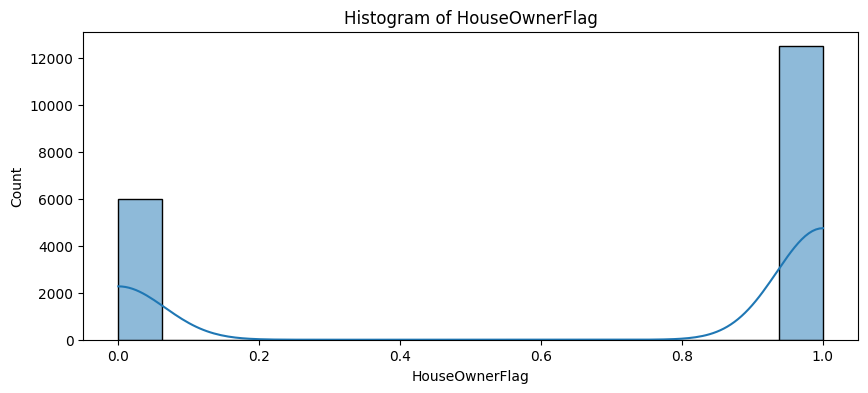
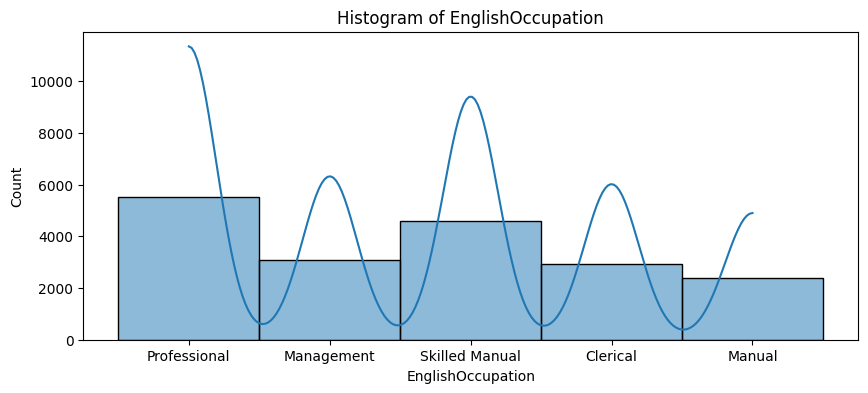
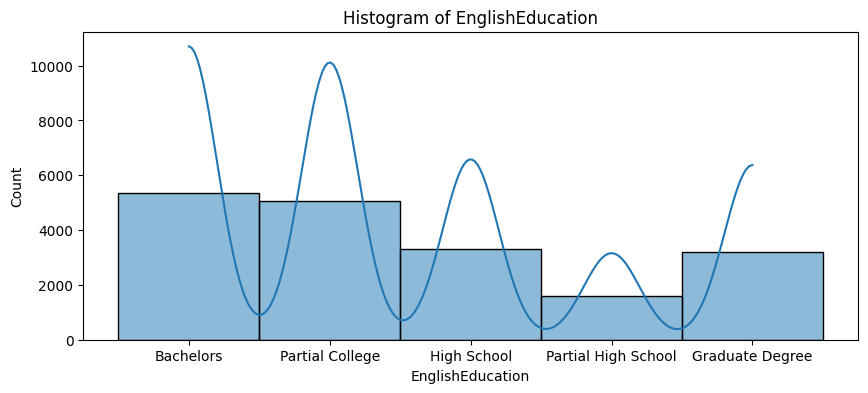
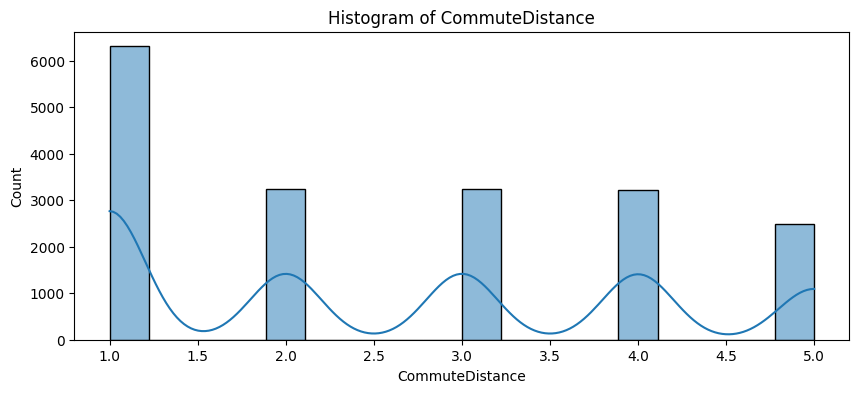
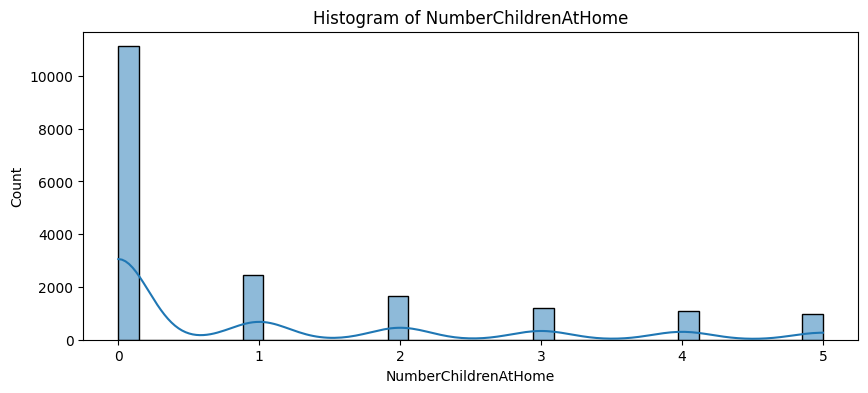
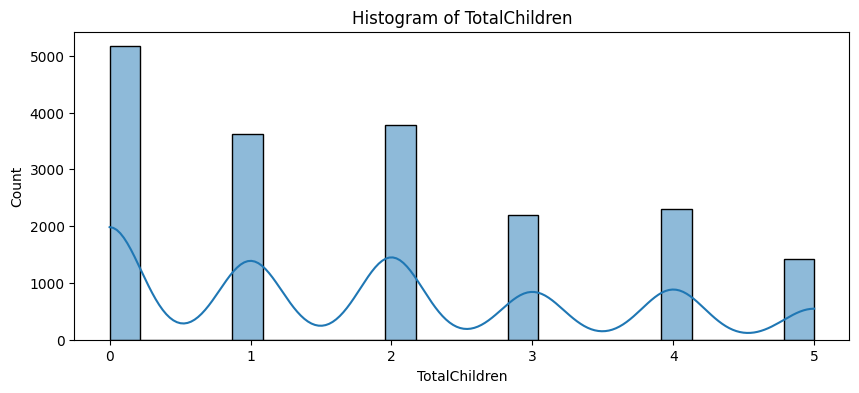
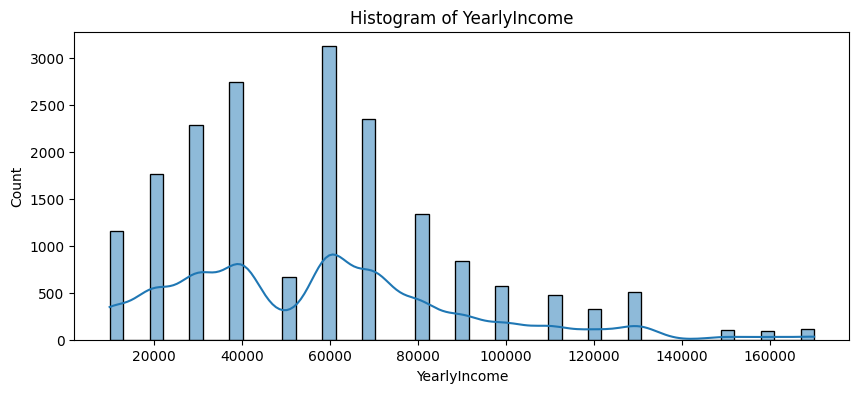
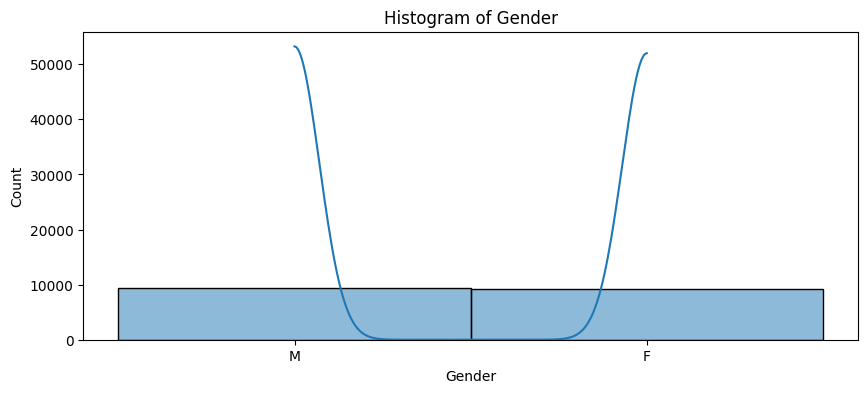
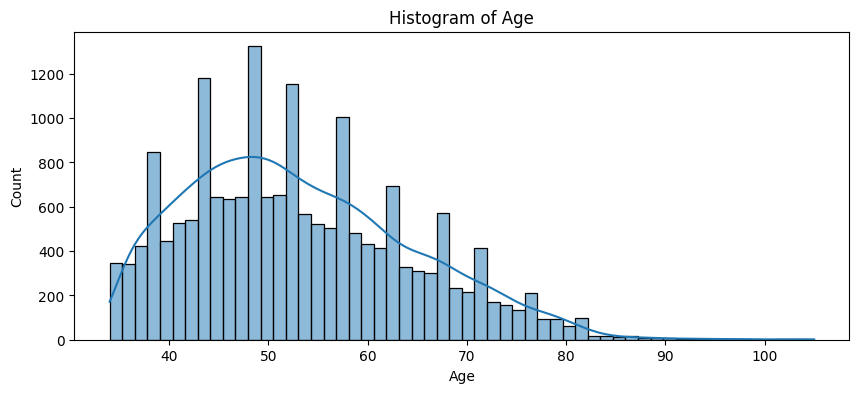
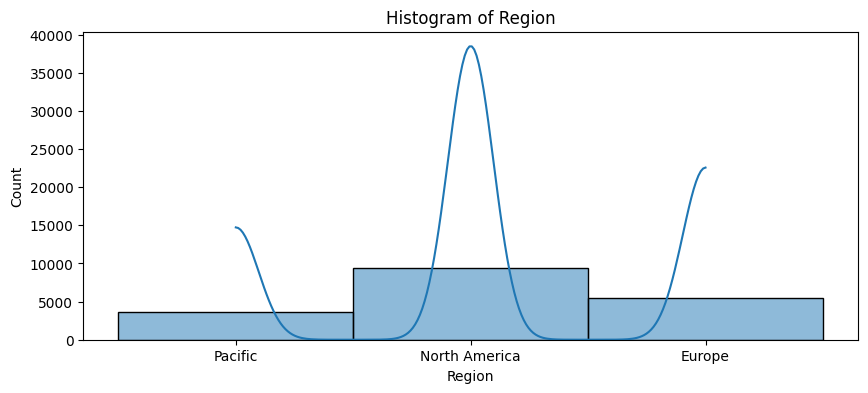
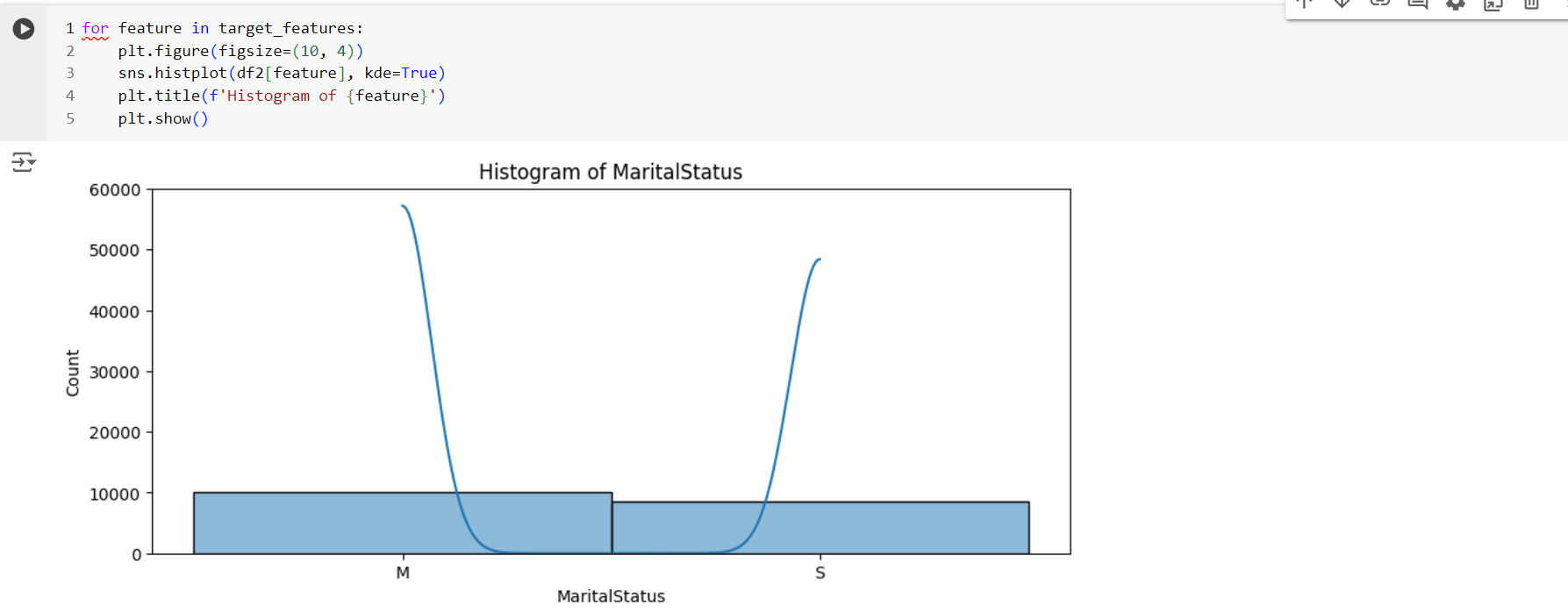
In order to perform mean/variance/standard deviation on Ordinal Numerica attribute which is CommuteDistance. We need to map the data of commute distance.  
Mapping is performed in the following order:  


After successfull mapping of CommuteDistance, we perform mean/variance/standard deviation on CommuteDistance

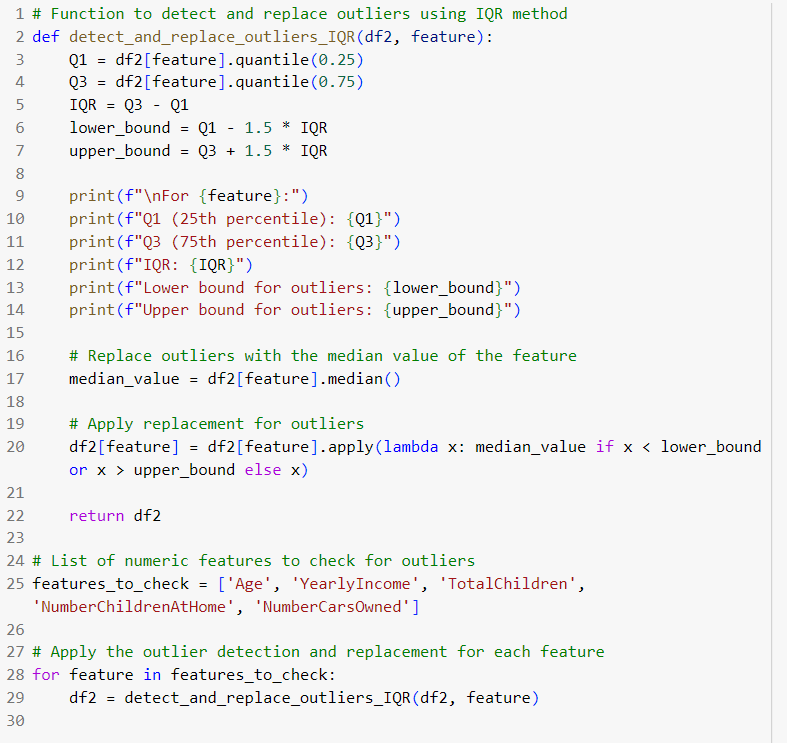


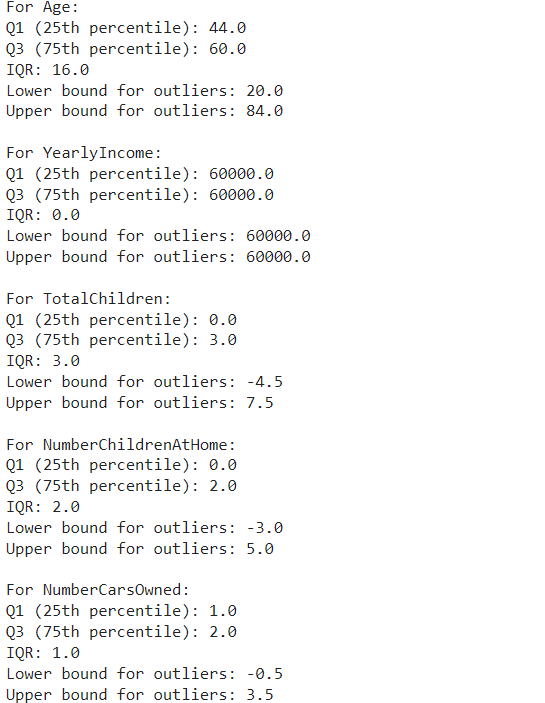
* Identify Data Dispersion (Distribution) for each feature using one of the methods: Box plot, Histogram, or z-score visualization

For performing Data Dispersion(Distribution) for each features using Histogram :  


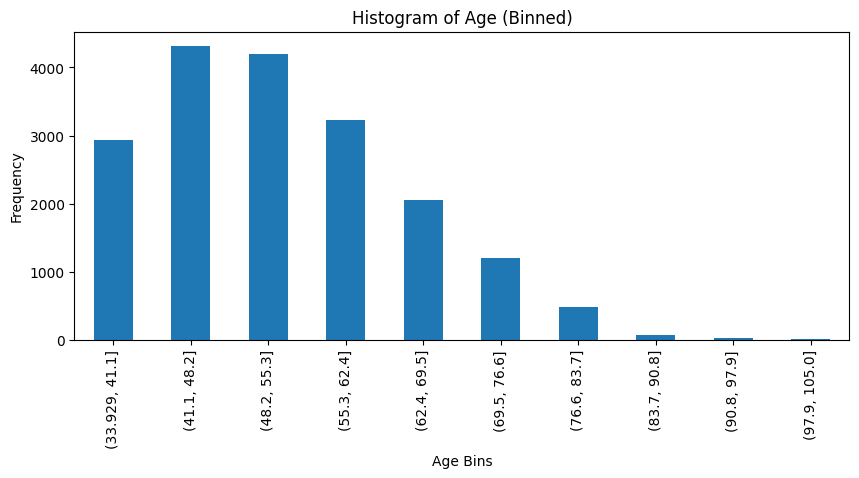
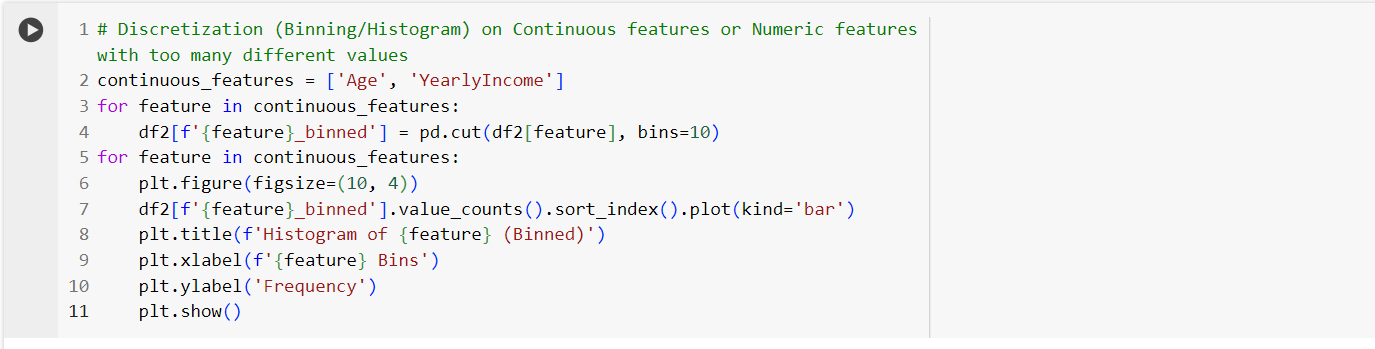


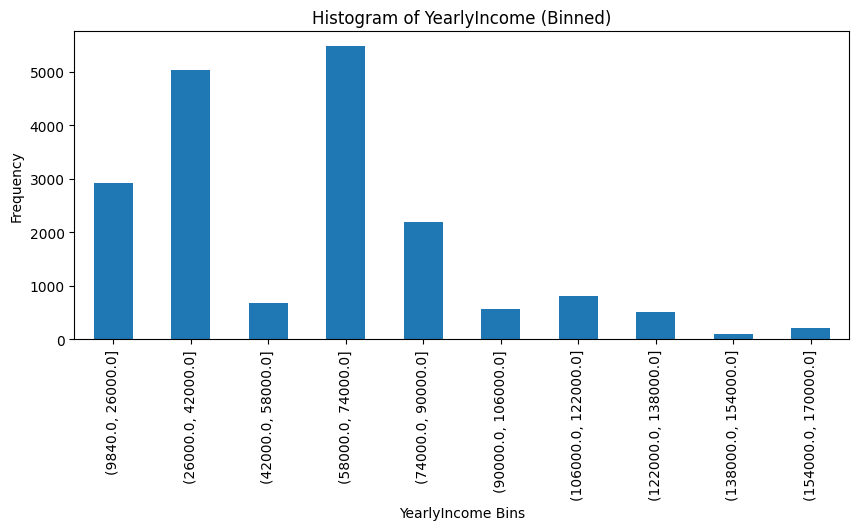
* Identify Outliers with 1.5IQR Method to Replace

As we check outliers for Numerical features,which are mentioned below in features\_to\_check:  


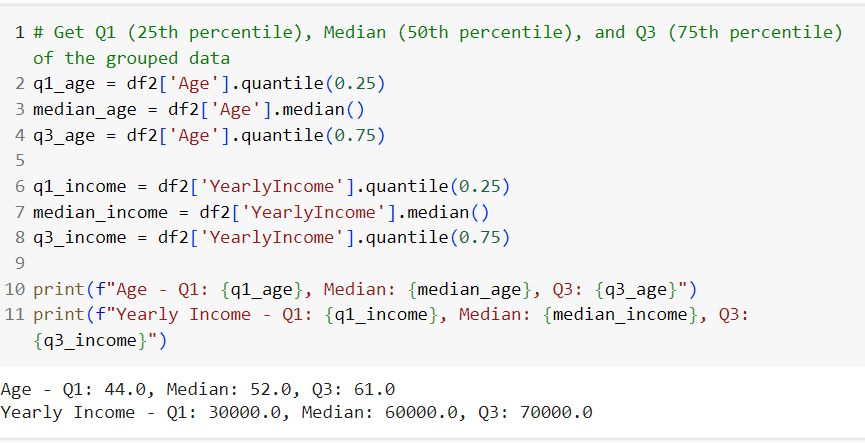
OUTPUT:  


* Discretization (Binning/Histogram) on Continuous features or Numeric features with too many different values



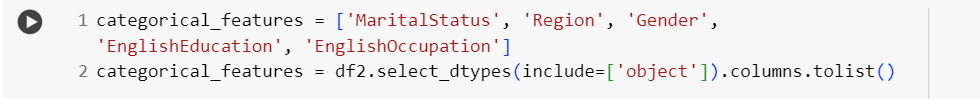
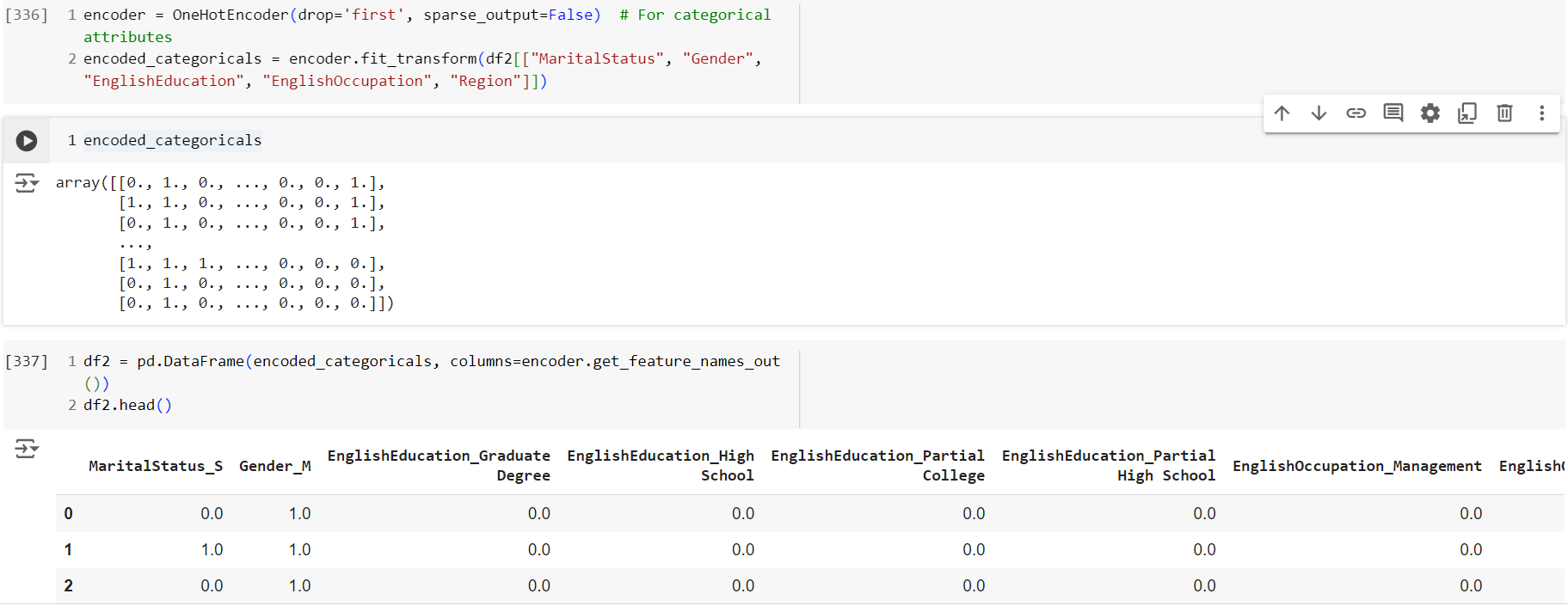
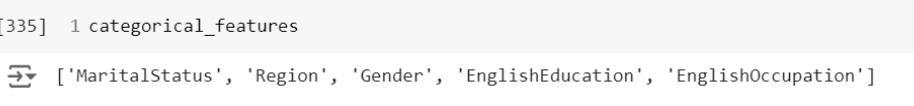


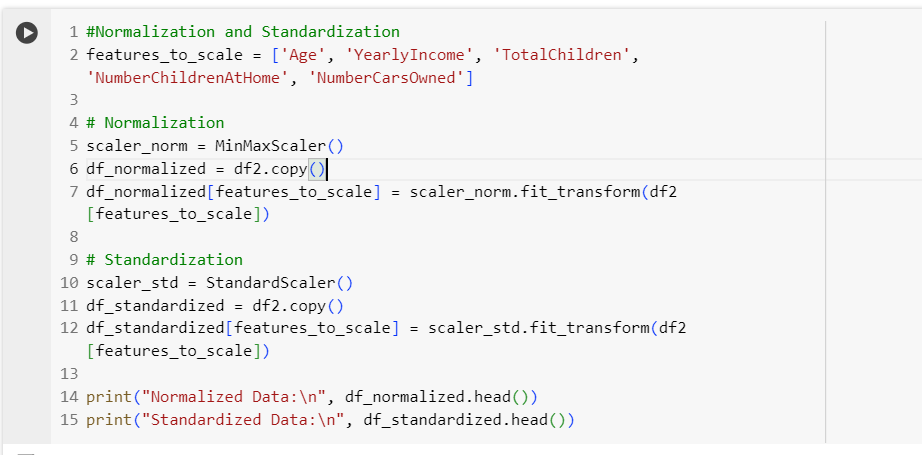
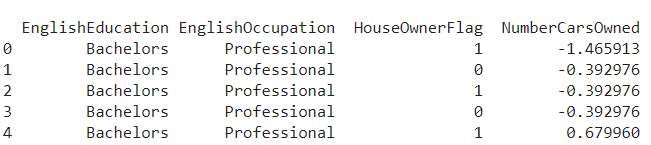
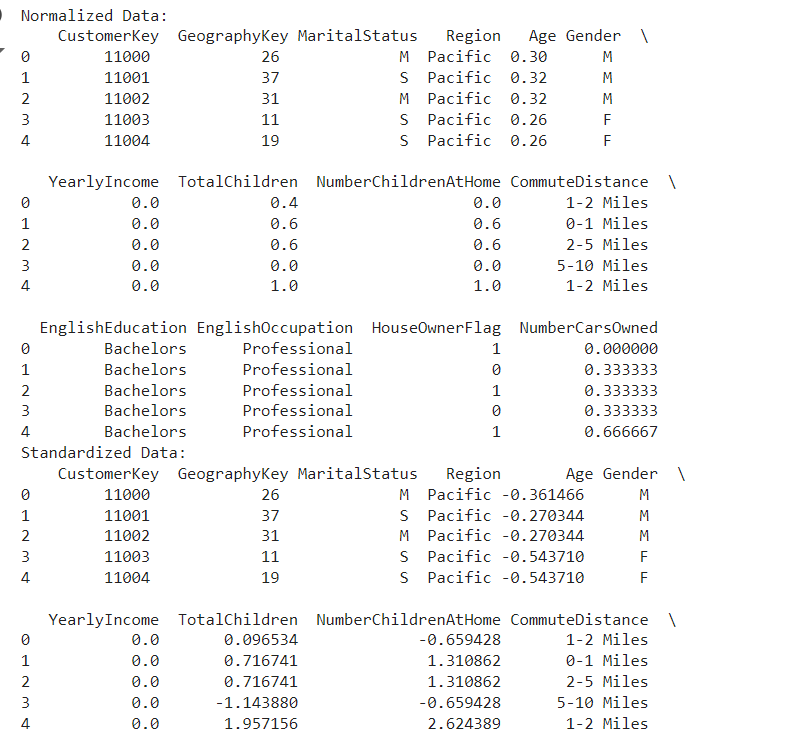
* Get Q1, Median, Q3 of the Grouped data (those features that were transformed into Histograms by Discretization)



* Identify which feature needs to be transformed by Binarization (One Hot Encoding) to apply Binarization for those features

We perform Binarization for the categorial features which are determined below:

* Normalization or Standardization  
  OUTPUT:  
  
* Saving the final processed data in a new csv file named as “VTargetBuyers\_processed.CSV  
  