**Matt Purvis – BUA 6110 – Predictive & Prescriptive Analytics – Bank Customer DDD**

1. **Please provide the specific list of predictors you identified for data exploration and due diligence activities. For each predictor, please show its data type (i.e., numeric or character variable)**

The list of predictors I have come up with to achieve the best results are:

|  |  |
| --- | --- |
| Variable | Data type |
| annual\_revenue | Numeric |
| MMDA\_flag | Character |
| Active\_user | Character |
| HELoans\_HELOC\_flag | Character |
| CDs\_flag | Character |
| Dwelling\_Type | Character |
| HH\_EDUCATION\_VALUE | Character |
| MARITAL\_STATUS | Character |
| ACTIVE\_USER | Character |

*\*Please note these variables showed a relationship with the ‘Customer Value’ field as part of EDA analysis and this does not reflect any inputs into the data model.*

1. **Are there any missing values? How’s the % of customers with missing value for each predictor. For predictor with missing values, describe and show how you would treat missing value**

|  |  |  |
| --- | --- | --- |
| Variable | # Missing | Action |
| annual\_revenue | 8 | Replaced the values with the median values of annual\_revenue |
| MOBILE\_ACTIVE\* | 2,997 | Boolean value: replaced nulls with 0, which equals ‘No’ |
| OLB\_ACTIVE\* | 1,311 | Boolean value: replaced nulls with 0, which equals ‘No’ |
| OLB\_ENROLLED | 644 | Boolean value: replaced nulls with 0, which equals ‘No’ |

*\*Note: The missing values were part of the 3rd party dataset but were not used due to a direct correlation with the ‘Active User’ variable.*

It is also worth noting that there were other missing variable actions taken. Please see the ‘.HTML’ file attached with my submission to see explanation of treatment of additional missing values.

1. **Are there any outliers? Show how you detect outliers and how you would treat outliers**

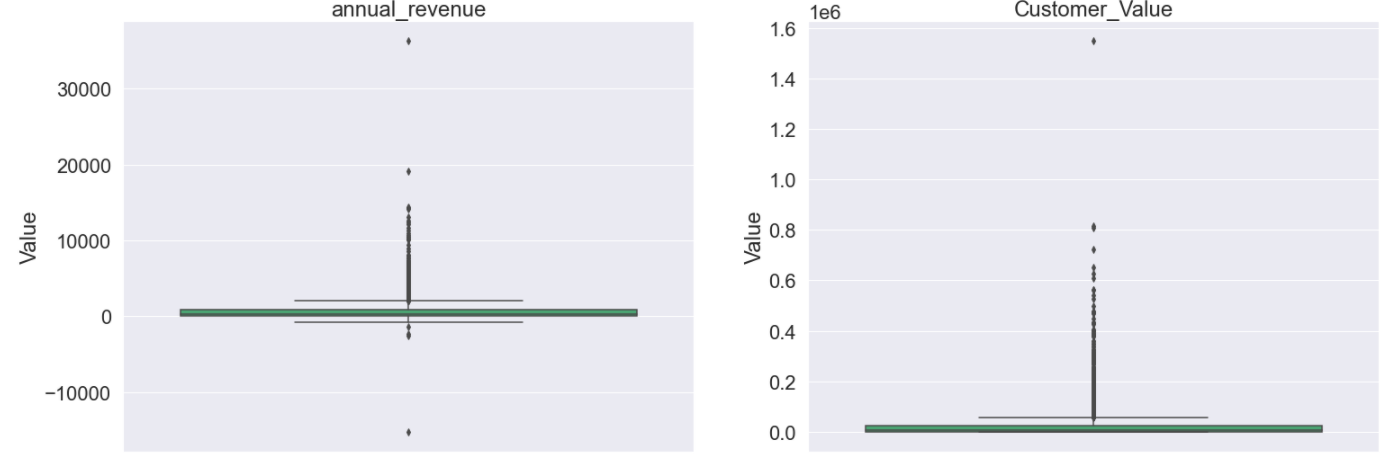
Outliers were assessed for both the ‘annual revenue’ and ‘Customer Value’ fields. First, I took the 99th and 1st percentiles for each variable and then I created a ceiling and floor to cap the high and low outliers.

Below is a summary of the action taken for outliers:

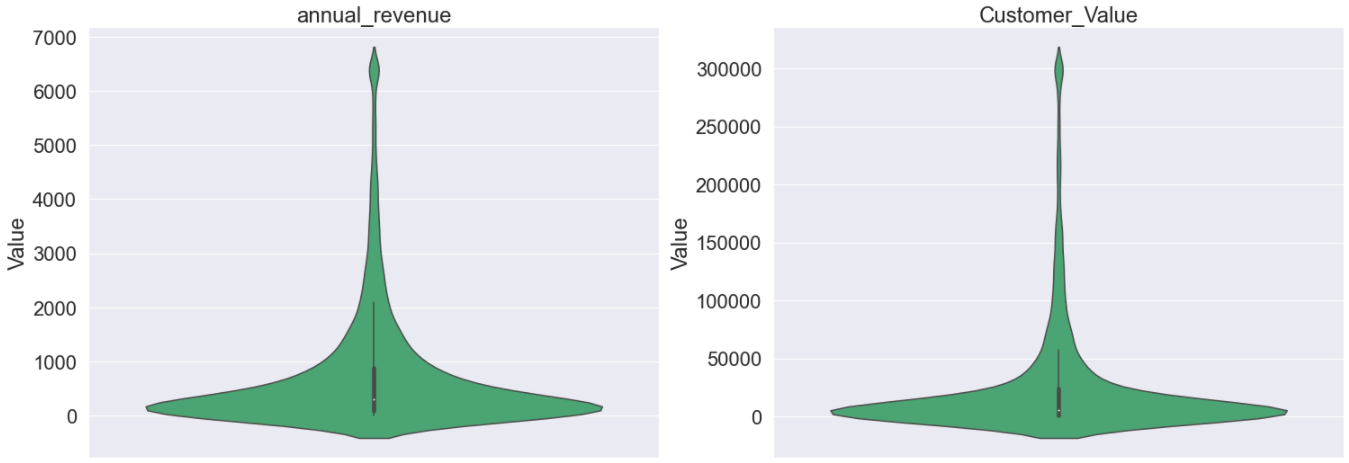
* Assess the 99th and 1st percentiles for each variable:
  + The 99th percentile in annual\_revenue is: 6374
  + The 1st percentile in annual\_revenue is: -40
  + The 99th percentile in Customer\_Value is: 296,261.
  + The 1st percentile in Customer\_Value is: 0
* Annual Revenue:
  + Ceiling: Use 99th percentile at 6400
  + Floor: Use 0 as the lowest allowable value
* Customer\_Value columns:
  + Ceiling: Use a number approximating the 99th percentile ($300,00)
  + Floor: Use 0 as the lowest allowable value

The following page shows the before and after. Please note that I use a box plot for the *before visual* and a violin plot for the *after visual*. The reason for this is to assess the final distributions of the numeric variables.

**Before:**



**After:**

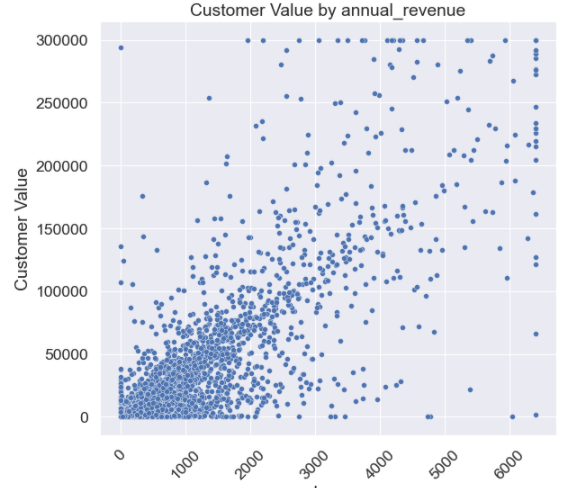


1. **Are there any character variables? If yes, recode character variables to numeric variables so you can use them for subsequent data exploration and model build**

Yes, there were several binary character variables that had either a 0 or 1 input for the value, which signified Yes (1) or No (0). For the Gender\_P1 column, Male = 0 and Female = 1.

1. **What are the relationship and strength of the relationship between each predictor and dependent variable? Show supporting data to answer this question**

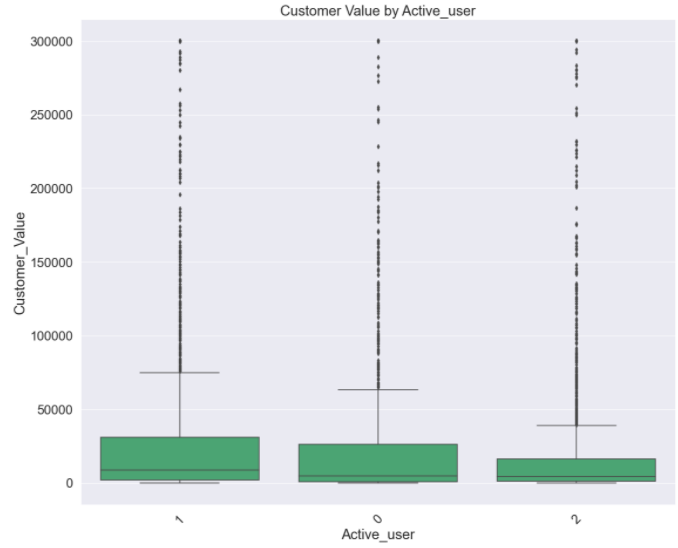
**Annual\_revenue:**



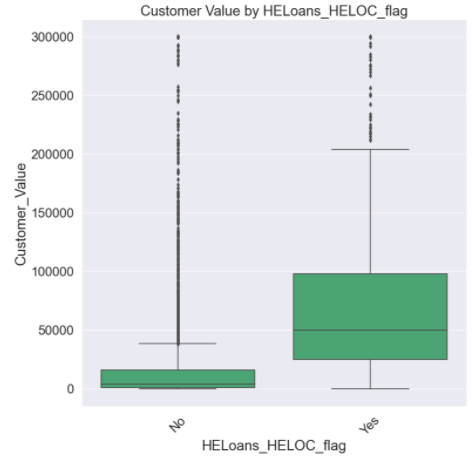
**MMDA\_flag**



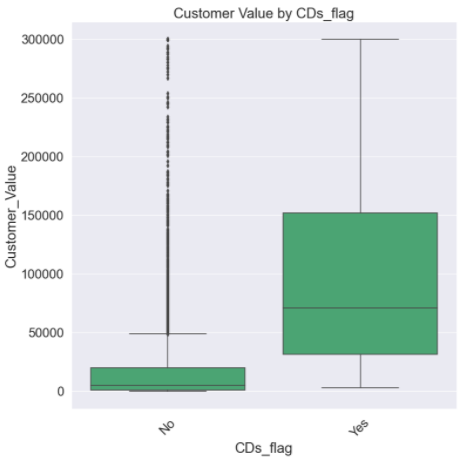
**Active\_user**



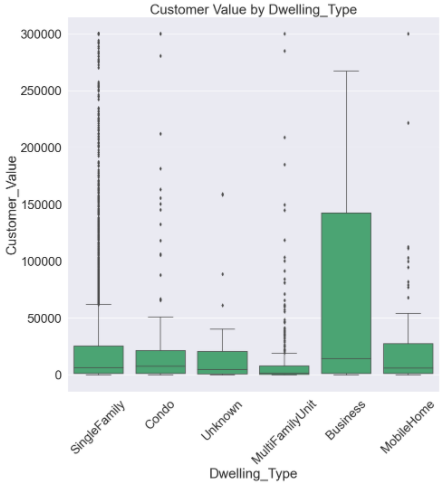
**HELoans\_HELOC\_flag**



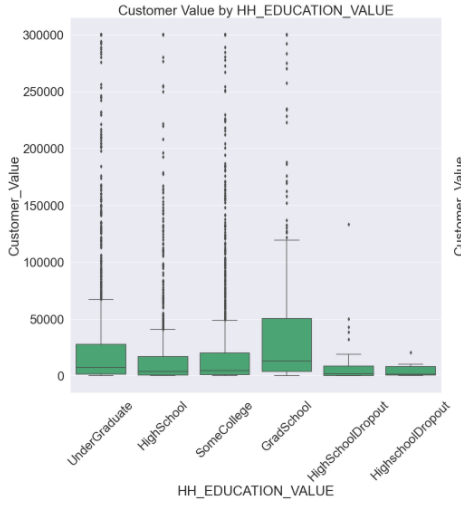
**CDs\_flag**



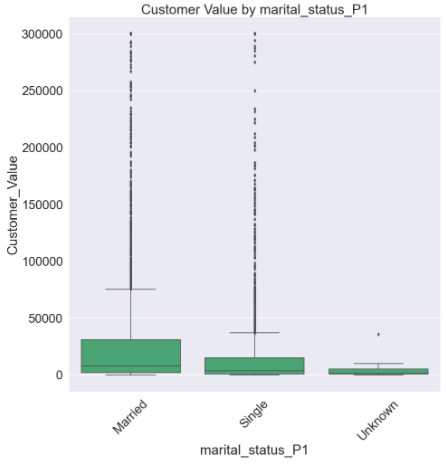
**Dwelling\_Type**



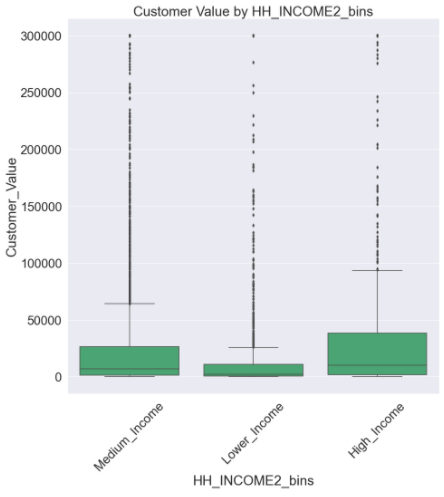
**HH\_EDUCATION\_VALUE**



**Marital\_status\_P1**



**HH\_INCOME2\_bins**



1. **For subsequent LINEAR regression model build, the relationship between each predictor and the dependent variable is assumed to be linear. Based on your findings from #5 above, if the relationship is not linear, show how you would recode / transform your predictors to linearize and strengthen the relationship to support a robust linear regression model build**

Each of the predictor variables seem to have some linear-like correlation.