**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**

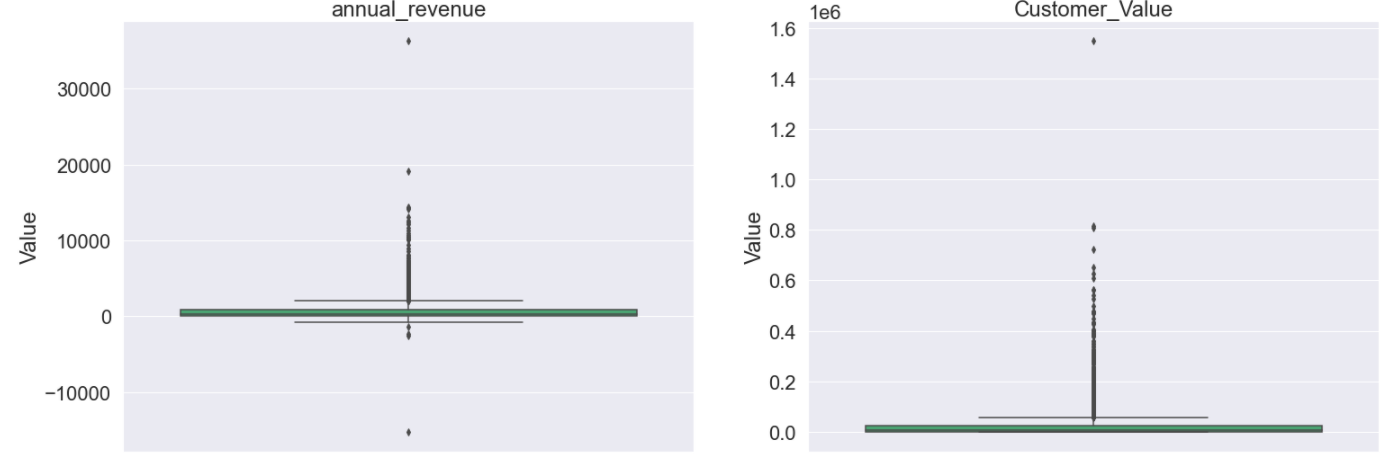
Outliers were assessed for both the ‘annual revenue’ and ‘Customer Value’ fields. First, I took the 99th and 1st percentiles for each variable and then I created a ceiling and floor to cap the high and low outliers.

Below is a summary of the action taken for outliers:

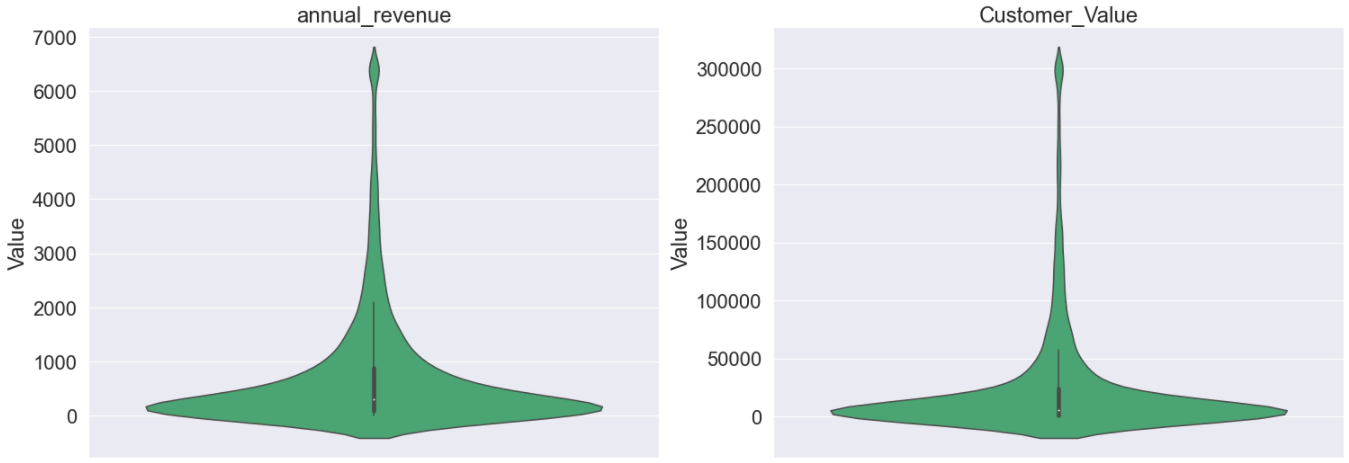
* Assess the 99th and 1st percentiles for each variable:
  + The 99th percentile in annual\_revenue is: 6374
  + The 1st percentile in annual\_revenue is: -40
  + The 99th percentile in Customer\_Value is: 296,261.
  + The 1st percentile in Customer\_Value is: 0
* Annual Revenue:
  + Ceiling: Use 99th 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:**

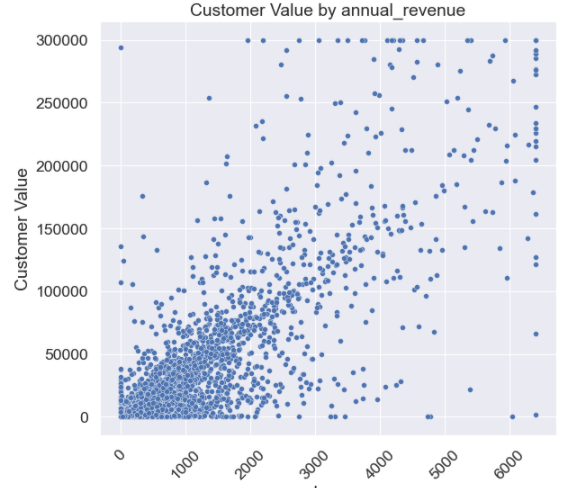


1. **Are there any character variables? If yes, recode character variables to numeric variables so you can use them for subsequent data exploration and model build**

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**

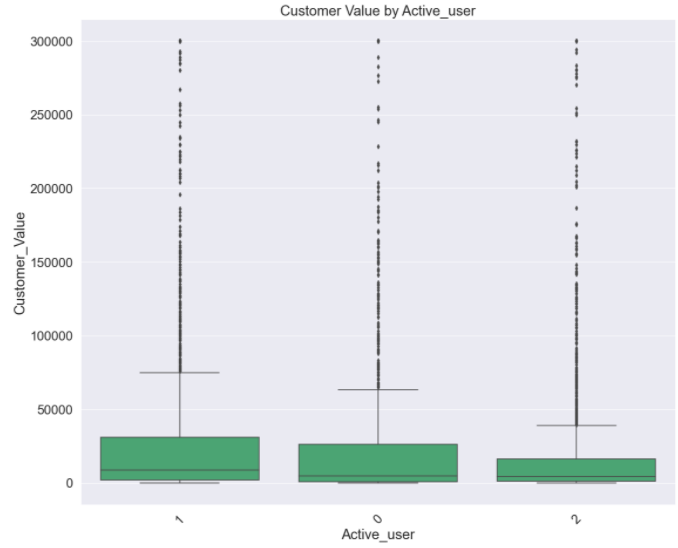
**Annual\_revenue:**



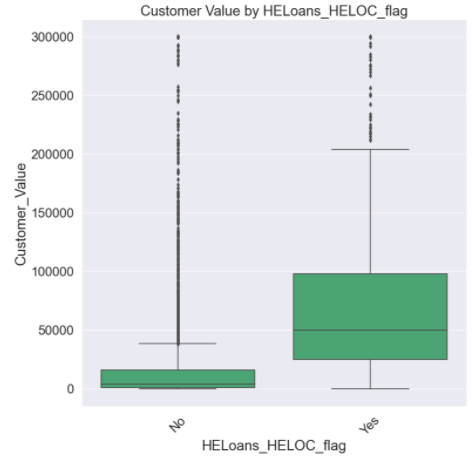
**MMDA\_flag**



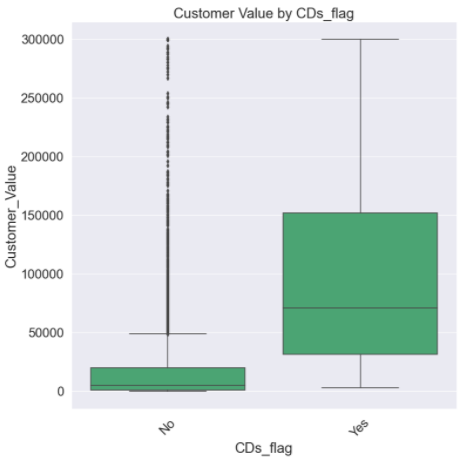
**Active\_user**



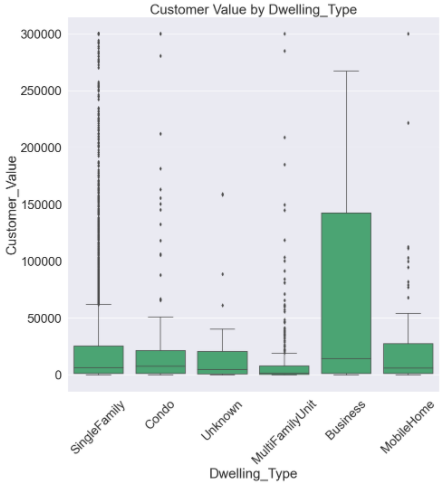
**HELoans\_HELOC\_flag**



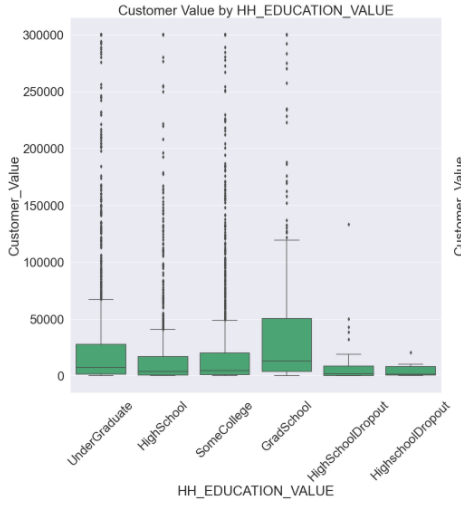
**CDs\_flag**



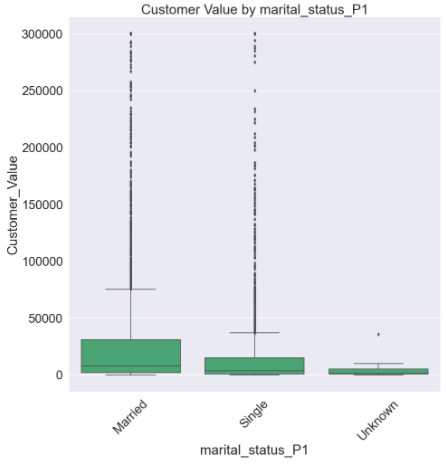
**Dwelling\_Type**



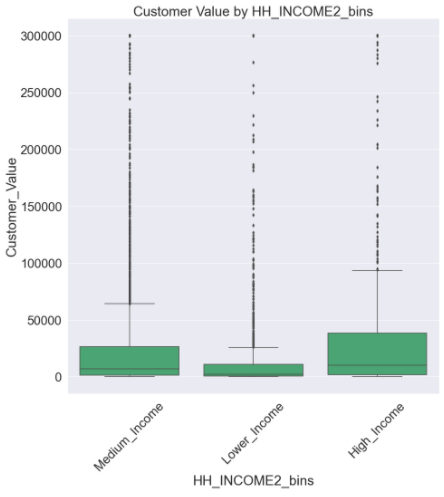
**HH\_EDUCATION\_VALUE**



**Marital\_status\_P1**



**HH\_INCOME2\_bins**



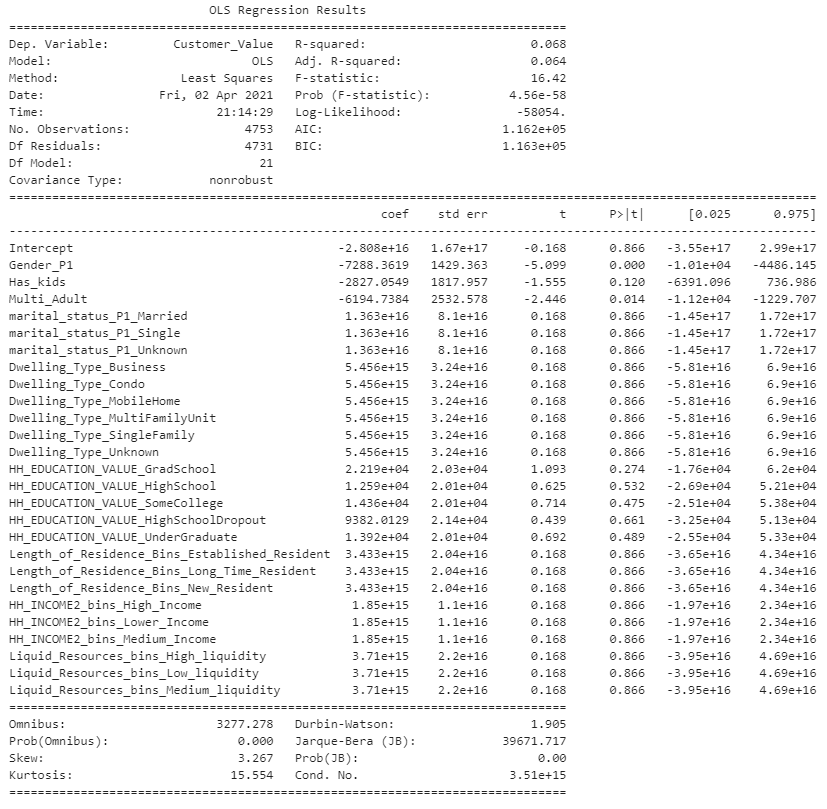
1. **For subsequent LINEAR regression model build, the relationship between each predictor and the dependent variable is assumed to be linear. Based on your findings from #5 above, if the relationship is not linear, show how you would recode / transform your predictors to linearize and strengthen the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

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

Question:

* Correlation of each predictor.. do we include all variables or only the variables we have assessed as the predictor variables? (see below)
* What are some example diagnostic plots you are looking for?

1. Show correlation of each predictor with DV (i.e., univariate relationship).



1. Show diagnostic plots of your predictors
2. For your initial multiple linear regression model –
   1. What are significant variables in the model in predicting customer value?
   2. What are variables that are not significant?
   3. Are any variables with opposite relationship with DV in the univariate correlation analysis vs. in your multiple linear regression model?
   4. What is the impact of significant variables?
   5. What is the goodness of fit (i.e., R-squared) of the model?
   6. Calculate predicted customer value for a particular customer using your model results
3. Now, remove variables that are not significant. Rerun your model and answer the same set of questions above.
4. Continue to tweak your model if needed, for example, including interaction terms, and finalize your model.
5. Comment on your final model results including
   1. Interpretation of the model
   2. Improvement of model performance from your initial model
   3. How will you use the model to support the Bank’s acquisition campaign?
   4. Any other comments you want to add

Deliverables –

1. Submit your report with supporting diagnostic and model outputs
2. Specifically, please include responses and comments to the questions specified for initial model as well as final model