Summary of the Data Set

Question 1 - Create a Histogram for a continuous variable

The variable considered for this analysis is MEDIAN_HOME_VALUE. This value is continuous in nature.
Provide the mean, median, standard deviation, and confidence intervals.

a) **Mean:** 1079.8719285566797 (in $100)

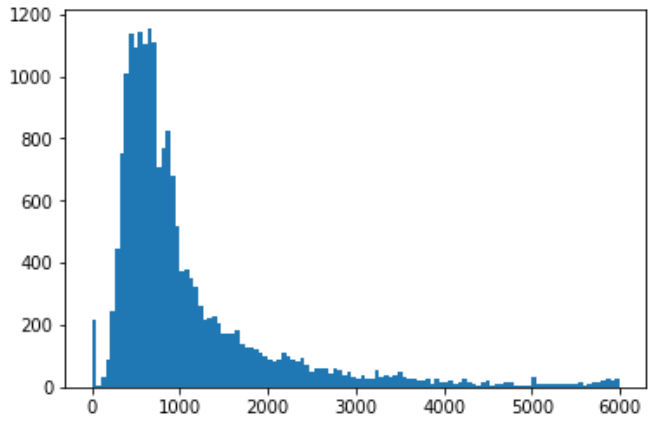
Median: 747 (in $100)

Std. Dev: 960.728

Confidence Interval: [1091.627, 1140.116] at 95%

b) Explain what these descriptive statistics tell us about the variable distribution
The median value is less than the mean value for the given description. Hence, it can be inferred that the distribution of this variable is highly right skewed. It indicates the presence of outliers in the data, may include some very wealthy donors which have a very high median home value. High standard deviation indicates that the variance or the spread of the data is also very high. Hence, the confidence intervals obtained will be large which is evident from the result.

The following histogram evidently shows the right skew.



c) Does the variable follow a normal distribution? Explain your answer.
The variable does not follow a normal distribution since the mean is not equal to the median. In this case the mean is greater than the median, this is a right skewed distribution. This has happened due some very wealthy donors which have a very high-priced home value.

Question 2 - **Create a Correlation matrix for all continuous variables and Chi-square test of association for all categorical variables**

a) Provide examples of the different ways' variables may or may not be correlated and explain.

MEDIAN_HOUSEHOLD_INCOME – MEDIAN_HOME_VALUE

Show a positive correlation. The explanation for this is generally people with higher incomes can afford costlier homes

LAST_GIFT_AMT – MEDIAN_HOUSHOLD_INCOME

Shows a positive correlation. The explanation is that people with higher income will donate generously

NUMBER_PROM_12 – MONTHS_SINCE_LAST_GIFT

Shows a negative correlation. The explanation is that if the number of promotions is high, then the gift received can be most recent. Hence, the more the number of promotions, the lesser the month since last gift.

b) Provide examples of the different ways' variables may or may not be associated and explain.

	TARGET_B	IN_HOUSE	URBANICITY	CLUSTER_CODE	HOME_OWNER	DONOR_GENDER	INCOME_GROUP	PUBLISHED_PHONE	WEALTH_RATING	PEP_STAR	RECENT_STAR_STATUS	recency_freq_status
TARGET_B	0	0	0.0002	0	0.0596	0.1738	0.0001	0.6542	0.0025	0	0	0
IN_HOUSE	0	0	0	0	0	0	0	0.2002	0	0	0	0
URBANICITY	0.0002	0	0	0	0	0	0	0	0	0	0	0
CLUSTER_CODE	0	0	0	0	0	0.0073	0	0	0	0	0	0
HOME_OWNER	0.0596	0	0	0	0	0	0	0	0	0.0936	0.0011	0
DONOR_GENDER	0.1738	0	0	0.0073	0	0	0	0	0.0454	0	0.0086	0
INCOME_GROUP	0.0001	0	0	0	0	0	0	0	0	0	0	0
PUBLISHED_PHONE	0.6542	0.2002	0	0	0	0	0	0	0	0	0.3638	0.0003
WEALTH_RATING	0.0025	0	0	0	0	0.0454	0	0	0	0	0	0
PEP_STAR	0	0	0	0	0.0936	0	0	0	0	0	0	0
RECENT_STAR_STATUS	0	0	0	0	0.0011	0.0086	0	0.3638	0	0	0	0
recency_freq_status	0	0	0	0	0	0	0	0.0003	0	0	0	0

We can see that there is a strong association between most of the categorical columns.

CLUSTER_CODE – URBANICITY –

There is a strong association between the URBANICITY and the CLUSTER_CODE as the clusters are decided by factors such as socioeconomic status, URBANICITY and other demographics. So, the variable URBANICITY will have a great influence in deciding the Cluster codes as similar type of cities will form the same cluster.

HOME_OWNER – WEALTH_RATING –

Wealth rating determines the capabilities of a person and determines the power of purchasing. Hence this association is justified.

PEP_STAR – HOME_OWNER –

The association between the PEP_STAR and HOME_OWNER can be justified as, if the person is a homeowner then there is a high probability that the person can be a pep star.

Question 3 Build a Linear Regression Model using a target and predictor variables

a) Provide detailed explanation of the results as it relates to the model fit, statistical measure of the variables, and model assumptions

Applying OLS model on the target variable TARGET_D and all other variables as predictor variables (except for MONTH_SINCE_FIRST_GIFT which had strong correlation 0.9 with MONTH_SINCE_ORIGIN), we find that following results:

OLS Regression Results

Dep. Variable:	TARGET_D	R-squared:	0.069
Model:	OLS	Adj. R-squared:	0.062
Method:	Least Squares	F-statistic:	9.299
Date:	Fri, 28 Jun 2019	Prob (F-statistic):	4.64e-139
Time:	15:20:07	Log-Likelihood:	-48944.
No. Observations:	13560	AIC:	9.811e+04
Df Residuals:	13451	BIC:	9.893e+04
Df Model:	108		
Covariance Type:	nonrobust		

R-Squared value gives the measure of fit of the model. It is the measure of how much the predictor variable can explain the target variable. We can see that the R^2 value is very low. This indicates that the model is not a correct fit.

The variables used for prediction can be identified as important or not important from the following values:

- coef
- std err
- t
- P>|t|

P value obtained from the output test the null hypothesis
$H_0 = \beta = 0$
$H_1 = \beta \neq 0$

If the p-value is greater than 0.05 at 95% confidence interval, then we cannot reject the null hypothesis and can conclude that the predictor variable used can be dropped for better results. Since, the output has 112 coefficients due to one hot encoding of the categorical columns. This is the extracted list of all the coefficients which came out significant (had p value < 0.05).

Col_Name	Col_coef	p_val
MEDIAN_HOME_VALUE	x3	4.95E-03
LAST_GIFT_AMT	x8	8.77E-120
MONTHS_SINCE_LAST_GIFT	x10	2.72E-02
URBANICITY_C	x12	2.41E-09
URBANICITY_R	x13	2.43E-09
URBANICITY_S	x14	5.44E-10
URBANICITY_T	x15	2.46E-02
URBANICITY_U	x16	9.54E-08
CLUSTER_CODE_13.0	x28	2.87E-02
CLUSTER_CODE_27.0	x42	1.18E-02
CLUSTER_CODE_28.0	x43	3.78E-04
CLUSTER_CODE_34.0	x49	5.43E-09
CLUSTER_CODE_35.0	x50	1.85E-16
CLUSTER_CODE_36.0	x51	6.24E-16
CLUSTER_CODE_37.0	x52	1.22E-11
CLUSTER_CODE_38.0	x53	7.25E-13
CLUSTER_CODE_39.0	x54	2.71E-16
CLUSTER_CODE_40.0	x55	4.46E-20
CLUSTER_CODE_41.0	x56	8.93E-12
CLUSTER_CODE_46.0	x61	2.56E-02
DONOR_GENDER_F	x70	1.64E-18
DONOR_GENDER_M	x71	1.93E-18
DONOR_GENDER_U	x72	2.49E-18
INCOME_GROUP_3.0	x74	1.88E-02
INCOME_GROUP_4.0	x75	1.07E-02
INCOME_GROUP_5.0	x76	5.79E-05
INCOME_GROUP_6.0	x77	2.08E-04
INCOME_GROUP_7.0	x78	1.35E-03
PEP_STAR_1	x89	3.39E-04
recency_freq_status_E2	x95	3.66E-03

Hence the important variables are:

- MEDIAN_HOME_VALUE
- LAST_GIFT_AMT
- MONTHS_SINCE_LAST_GIFT
- DONOR_GENDER
- PEP_STAR

There exists a strong multicollinearity in the dataset. This maybe because the proportion of cases in the reference category is small, even if the categorical variable is not associated with other variables in the regression model. Suppose, for example, that a marital status variable has three categories: currently married, never married, and formerly married. You choose formerly married as the reference category, with indicator variables for the other two. What happens is that the correlation between those two indicators gets more negative as the fraction of people in the reference category gets smaller. For example, if 45 percent of people are never married, 45 percent are married, and 10 percent are formerly married, the VIFs for the married and never-married indicators will be at least 3.0.

Assumptions of OLS Model:

1. x and y have a linear dependence
2. E(e) = 0 The error term should have zero mean
3. $Cov(e_i,e_j) = 0$ - The covariance between any pair of random errors, e_i and e_j is 0
4. Var(e) = c – the variance of e should be constant. Otherwise it leads to Heteroskedasticity causing large, non-robust standard errors in the model.

b) Provide suggestions to either improve the model's prediction or provide a new approach to the prediction.
The most important thing to look for improving the model predictions are:

Treating the skewness of the data – in the given dataset we have seen that that there are many columns including the TARGET_D column that is highly skewed. This causes the model to be asymmetric and hence the prediction results are off. The most important treatment to be done is by either taking a square root transformation or a log transformation of the variables which are skewed and have come up as important.

Treating Multicollinearity – As explained earlier there is a high possibility that the categorical predictors that have class imbalance due to which during the one hot encoding gives multicollinearity between columns causing the formation of a singular matrix.

Extreme data points – some of the data points may be extreme outliers which also cause the skewness in the results these have to be removed or capped to mean values to increase the accuracy.

Non-Linear models – It is safe to assume that the target variable and the predictor variables do not have a linear dependence. Hence, instead of applying linear methods other regression techniques such as kNN, SVR, Decision Tree and Random Forest regressors.

Question 4 - Compare the Linear Regression Model versus other Machine Learning Methods

a) **Train the same data and target with 3 other machine learning methods of your choice**

- kNN Regressor
- LASSO - Along with shrinking coefficients, lasso performs feature selection as well. It makes some of the coefficients zero
- Decision Tree Regressor

b) **Compare all models based on their results. Which is the best model?**

Model name	Model parameter	Train accuracy	Test accuracy
LR		0.06828	-0.141831
knnr	k = 15	0.089761	-0.048695
Lasso	alpha = 0.005	0.061611	-0.102223
DecisionTree		1	-3.508434

Out of the three models that were used other than the Linear model, kNN regressor performed the best. This can be seen by comparing the test and the train score.

c) **Explain which model statistics support the best model and why?**

The model statistic used for to select the best model is both Train and Test Score. A high train score indicates that the model has learnt well. And a high-test score indicated that the model has a good ability to predict on unseen data. According to the results, the train score is highest for the decision tree model. But it has a the worst test score. This indicates that there is a very high over fit in the model. In the remaining two models LASSO and kNN, kNN has both high test and train score. Hence, this is selected.

Question 5 - Build the Best Classification Model using Machine Learning Methods

a) **Train the same data using the binary target with at least 5 different machine learning methods of your choice**
Models used are:

1. Logistic Regression
2. kNN Classifier
3. Linear SVC
4. Decision Tree
5. Random Forest

b) **How would you go about comparing the accuracy of each?**
Instead of comparing the accuracy metric of the model, due to a strong class imbalance between the donor and the non-donor class, we will have to investigate the precision, recall and F1 values for the models. From the classification report generated for each model we would look at the Recall score for the Donor class (as it is the minority class) to select the best model.

Improving the recall of the donor class is important because the data has a lot of non-donor class. The models will learn better about that class. The learning however for the donor class will be poor. Hence there is a high possibility that False negatives in the predictions will increase. In other words we will start marking potential donors as non-donors and stop targeting them, ultimately leading to churn of such donors. Hence, whichever model reduces this, ultimately reduces the churn rate. So, whichever model will have the highest recall for Donor class is of importance.

c) **Which method provides the best accuracy based on the comparison? What were the accuracy measures you used to support the best the model?**
Based on the results, Decision tree model has a high recall for the donor class. When comparing the over all precision recall values the DT has comparable results to that of an RF model. Hence, we can go ahead with the Decision Tree model for classification.

d) **Explain why you think that method performed the best?**
The decision tree model tends to overfit. It is a complex model and hence it tends to perform the best over smaller sets of data. However, for regularization standpoint this would not hold good.

A more elegant solution would be to treat the data set for class imbalance by using sampling techniques, SMOTE and other such methods and making it more robust. Once, this is done an Ensemble technique such as Random Forest can be applied to the data set which will make it more robust for larger test sets. From the results it is evident that on an average RF also has a comparable result to that of DT and if the above-mentioned principles are applied that would be the perfect model.