Data Quality Plan

Feature	Data Quality Issue	Handling Strategy
age	Missing Values (31.3%)	Imputation - replace missing values with median
occupation	Missing Values (74.5%)	Drop feature
regionType	Missing Values (49.2%)	Drop feature
currentHandsetPrice	Missing Values (58.3%)	Drop feature
marriageStatus	Missing Values (39.6%)	Drop feature
handsetAge	Negative min des not make sense	Drop rows if number of -ve values small
creditCard	Irregular Cardinality	Rename values
numHandsets	Outliers	Do nothing
handsetAge	Outliers	Do nothing
avgBill	Outliers	Do nothing
avgMins	Outliers	Do nothing
avgrecurringCharge	Outliers	Do nothing
avgOverBundleMins	Outliers	Do nothing
avgRoamCalls	Outliers	Do nothing
callMinutesChangePct	Outliers	Do nothing
billAmountChangePct	Outliers	Do nothing
avgReceivedMins	Outliers	Do nothing
avgOutCalls	Outliers	Do nothing
avgInCalls	Outliers	Do nothing
peakOffPeakRatioChangePct	Outliers	Do nothing
avgDroppedCalls	Outliers	Do nothing
lifeTime	Outliers	Do nothing
lastMonthCustomerCareCalls	Outliers	Do nothing
numRetentionCalls	Outliers	Do nothing
numRetentionOffersAccepted	Outliers	Do nothing
newFrequentNumbers	Outliers	Do nothing

The above table shows the features which have data quality issues, identifies the issue and describes the proposed solution. A more detailed discussion of the decisions that were made is below.

Issues and Proposed Solutions

age contained a large amount of 0 values which did not make sense, since the age of a customer could not be 0. In the case of age I believe imputation by replacing these 0s with the mean makes sense as only 31.3% of values are missing and age may be an important feature for predicting churn.

occupation and regionType had 74.5% and 49.2% missing values respectively and as such I decided to drop these features. I did not think imputation with the mode would make sense since such a large number of features were missing, the data would no longer be representative of reality if this were done.

currentHandsetPrice had a large number of 0 values which did not make sense, because even a cheap phone has some value. I decided to drop the feature instead of using imputation because replacing the 58.3% missing 0s would have dragged up the mean and median significantly and created an unrealistic representation of customer handset price.

marriageStatus had 39.6% of values listed as unknown and I decided to drop this feature because of the amount of missing data. I did not think imputation would make sense because of the low cardinality of this feature. Replacing the missing values with either no or yes would unfairly represent one over the other in the data.

customer is an ID feature and would not provide any predictive power for determining the likelihood of a customer to churn so I decided to drop it.

creditCard had a cardinality of six despite representing a boolean feature. This was due to true and false values being recorded using different spellings. I decided to address this by replacing all falsey values with false and all truthy values with true.

The other features in the table contained outliers however, none of the features have outliers which are completely unreasonable and it is impossible to verify the validitiy of the extreme values without conferring with those who collected the data. Therefore I have decided to do nothing and to accept the values as valid and keep them in the dataset.