# Data Exploration

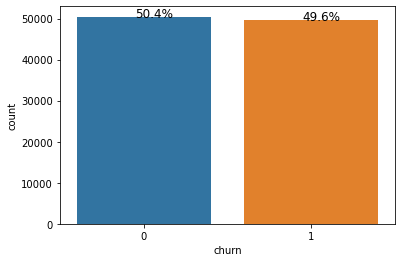
The following document provides a breakdown of the data for telecom churn customers wrt various fields

## Structure of the Data

|  |  |  |  |
| --- | --- | --- | --- |
|  | A | B | C |
| 1 | Data Type | Number of Columns | Number of Rows |
| 2 | Numerical Columns | 79 | 100,000 |
| 3 | Categorical Columns | 21 | 100,000 |
| 4 | Total | 100 | 100,000 |
| 5 |  |  |  |
| 6 |  |  |  |

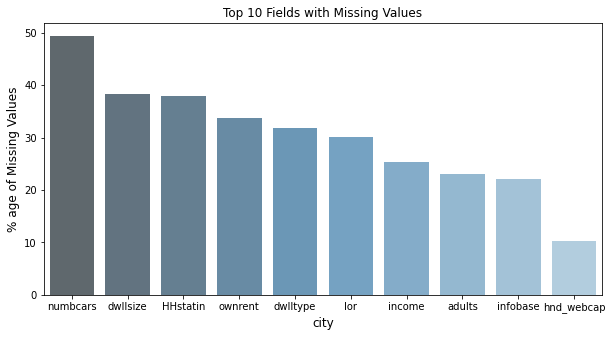
## Distribution of Churn and Not Churned Customers

This is an example of balanced dataset with equal proportion of churn and not churned customers.



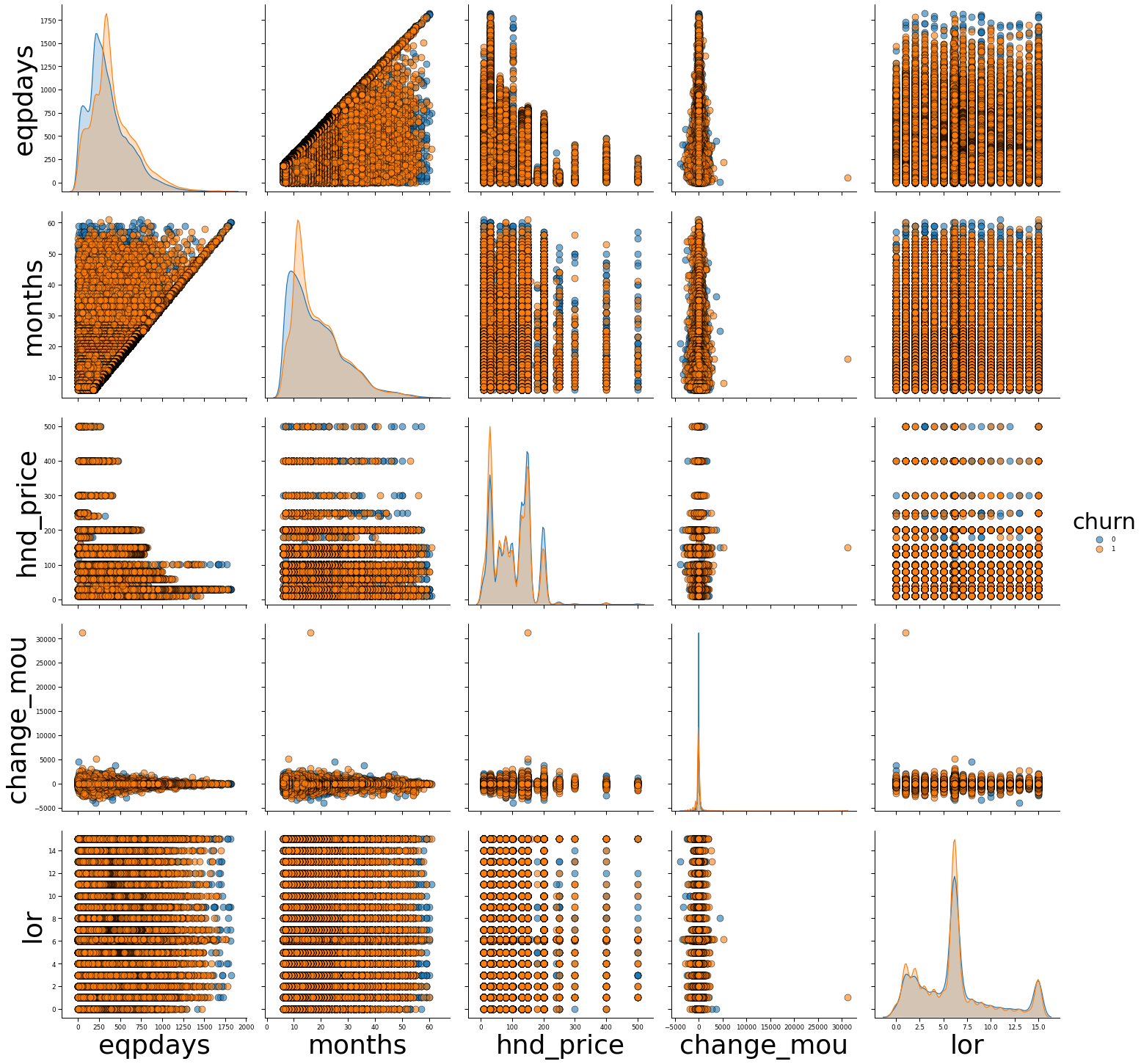
EDA Figure 1

## Missing Values in the columns



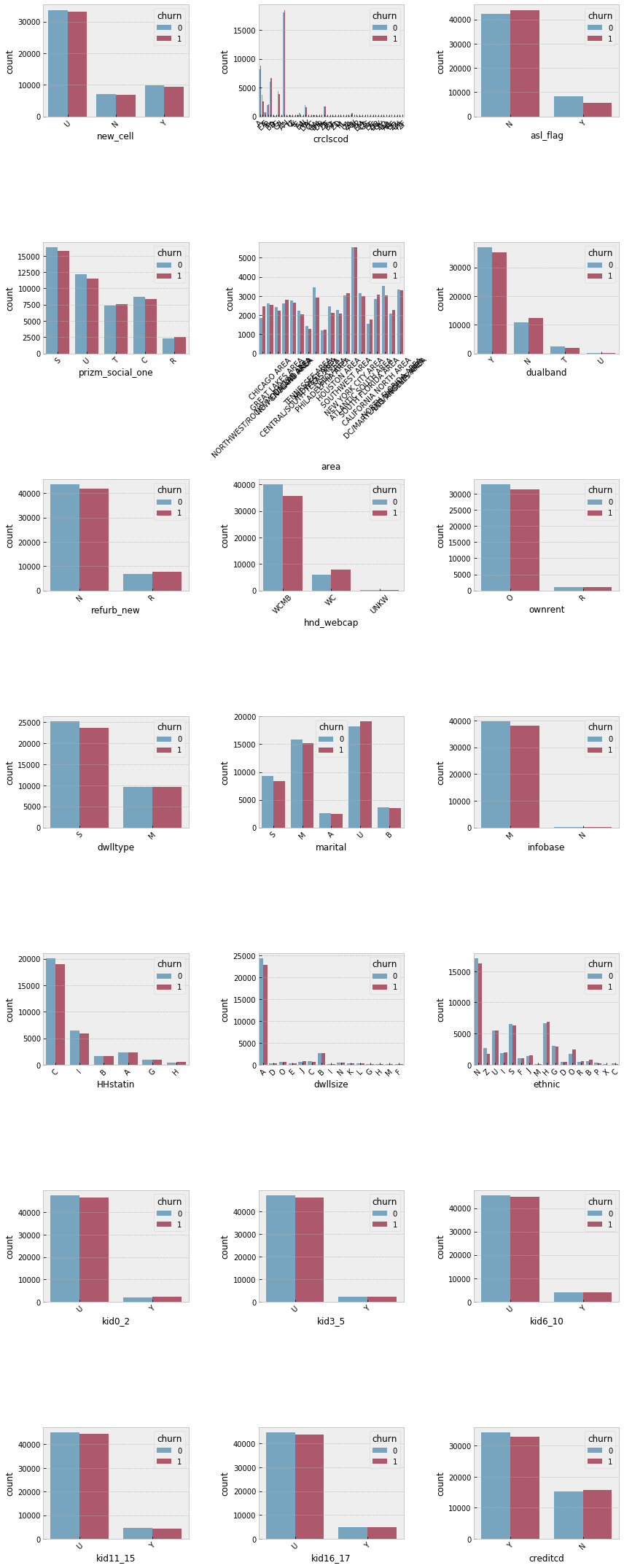
EDA Figure 2

## Correlation of Relevant Numerical Values with Churn



EDA Figure 3

## Relationship of Categorical Values with Churn



EDA Figure 4

**Observations**

**Churn :** From EDA Figure 1 , we can ascertain that the dataset is a balanced one with ~50% of customers churning and the remaining not churning. This implies that we don’t have to employ any sampling technique ( undersampling or oversampling) to drive performance gains

**Missing Values :** As per EDA Figure 2, it can be surmised that the most of the missing values is in the Personal customer attributes like number of cars, dwelling size, length of residence etc. Among the top 10 variables with missing values, 3 are numerical variables while the rest are categorical. Usually

Personal information of the customers are bought from 3P data providers by telecom companies. As a result, most of the data might be incomplete due to the usage of sparse and disparate data sources. So in order to proceed with model building, we would need to replace the missing values.

Different ways can be explored while imputing or replacing the missing values in the fields.

For Numerical variables, simply imputing the missing values with the corresponding mean of the field can drive performance gain significantly. Another method of imputation would be to use a simple or iterative imputer. Similarly for categorical variables , a simple replacement with the mode would suffice and help in improving performance. Additionally, using package based imputers might also help in imputing values.

**Numerical Variables :**

From EDA Figure 3 , it can be seen that the age of current equipment and number of months of service are somewhat normally distributed. Additionally, we can see that there is a linear relationship between the the age of equipment and months in service which is self explanatory.

**Categorical Variables :** Looking at the categorical variable’s relationship with churn , it is evident that the proportion of churned customers is equal to not churned customers for each value of the categorical variables. This confirms the hypothesis that there is no special concentration of churned customers across categorical values.

## Appendix

**Data Dictionary**

Numerical Variables



Categorical Variables



References

1. Matplotlib: <https://matplotlib.org/users/whats_new.html#figure-and-axes-creation-management>
2. Seaborn : <https://seaborn.pydata.org/tutorial/distributions.html>
3. Simple Imputer : <https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html>
4. Iterative Imputer : <https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html>
5. Feature Importance: <https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html>