Data Wrangling

Exploratory Analysis:

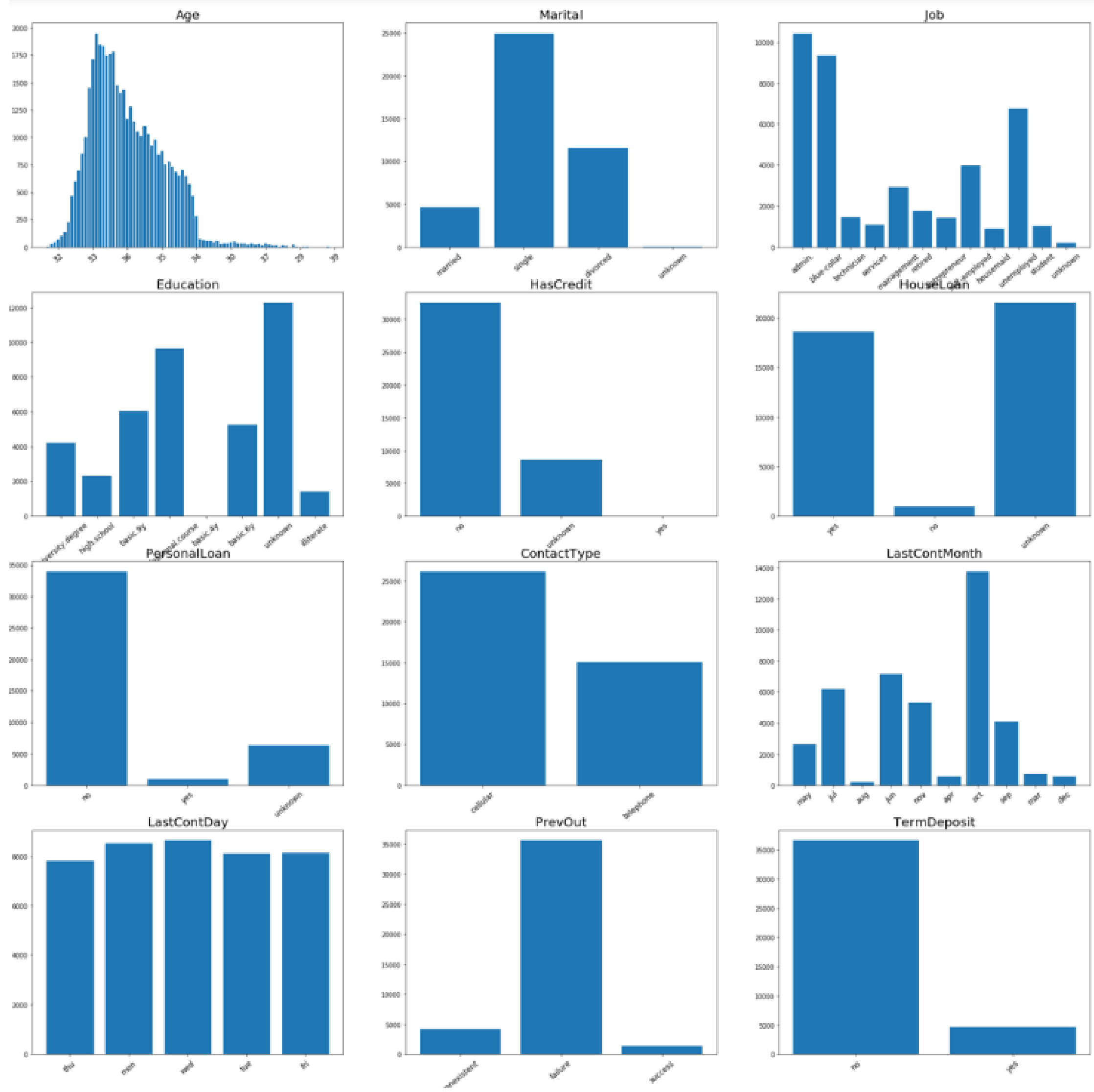
We first start the exploratory analysis of the missing or Unknown values of the categorical variables which are use for predicting the outcome Term Deposit. After loading the data we rename the columns and check the categorical variables (Job, Marital, Education, HasCredit, HouseLoan, PersonalLoan, ContactType, LastContMonth, LastContDay, PrevOut, TermDeposit) for missing values.

We will draw a subplot for these categorical variables for initial analysis:

```
categorical_variables = ['Age', 'Marital', 'Job', 'Education', 'HasCredit', 'HouseLoan', 'PersonalLoan', 'ContactType', 'LastConti
nrows=4
ncols=int(len(categorical_variables)/nrows)
fix,ax2d=plt.subplots(nrows,ncols, figsize=(30,30))

fig.subplots_adjust(wspace=0.6, hspace=6.0)
ax=np.ravel(ax2d)

for count,col in enumerate(categorical_variables):
    ax[count].bar(df[col].value_counts().index,df[col].value_counts().values)
    ax[count].legend(loc=1)
    ax[count].set_title(col, fontsize= 20)
    ax[count].set_xticklabels(df[col].value_counts().index, rotation=40, fontsize= 12)
plt.show()
```



There are unknown values for many variables in the Data set. One way to handle is to discard the row but that would lead to reduction of data set which wouldn't serve the purpose of building accurate and realistic prediction model. Another way is to infer the value from other variables, however it doesn't guarantee that all the missing values will be address but majority of them will be cleaned up for analysis. We will start cleaning data by updating the unknown values to Nan.

```
categorical_variables = ['Marital', 'Job', 'Education', 'HasCredit', 'HouseLoan', 'PersonalLoan', 'ContactType', 'LastContMonth',
for col in categorical_variables:
    df.ix[df[col]=='unknown',col] = np.nan
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
Age                 41188 non-null int64
Job                 40858 non-null object
Marital             41108 non-null object
Education           39457 non-null object
HasCredit           32591 non-null object
HouseLoan           40198 non-null object
PersonalLoan        40198 non-null object
ContactType         41188 non-null object
LastContMonth       41188 non-null object
LastContDay         41188 non-null object
LastContDuration    41188 non-null int64
Campaign            41188 non-null int64
PreviousDay         41188 non-null int64
PrevContNum         41188 non-null int64
PrevOut             41188 non-null object
EmpVarRate          41188 non-null float64
ConsumerPriceIdx    41188 non-null float64
ConsConfIdx         41188 non-null float64
Euribor             41188 non-null float64
Employeeno          41188 non-null float64
TermDeposit         41188 non-null object
dtypes: float64(5), int64(5), object(11)
memory usage: 6.6+ MB
```

Variables with Nan values are : Education, Job, HasCredit, HouseLoan, PersonalLoan and Marital.
However the Marital status has few unknown values, significant ones are Education, Job, HouseLoan and PersonalLoan. We will try to see the pattern for these missing values.

We will write a function which will return the Variable 1 groupby Variable 2 unique value counts as DataFrame.
We will use this function to check the missing/unknown values and see if can draw ay intuitions in filling the Nan values.

```python
def var_test(df,f1,f2):
    var1 = list(df[f1].unique())
    var2 = list(df[f2].unique())
    dataframes = []
    for e in var2:
        dfv2 = df[df[f2]==e]
        dfv1 = dfv2.groupby(f1).count()[f2]
        dataframes.append(dfv1)
    xx=pd.concat(dataframes, axis=1)
    xx.columns=var2
    xx=xx.fillna(0)
    return dfv2
```

```python
var_test(df,'Job','Education')
```

	basic.4y	high.school	basic.6y	basic.9y	professional.course	nan	university.degree	illiterate
admin.	77	3329	151	499	363	0.0	5753	1.0
blue-collar	2366	878	1448	3654	453	0.0	94	8.0
entrepreneur	137	234	71	210	135	0.0	610	2.0
housemaid	516	174	77	94	59	0.0	139	1.0
management	100	298	85	166	89	0.0	2186	0.0
retired	601	276	75	145	243	0.0	286	3.0
self-employed	93	118	25	220	168	0.0	765	3.0
services	132	2832	226	388	218	0.0	173	0.0
student	26	357	13	99	43	0.0	170	0.0
technician	58	873	87	384	3330	0.0	1809	0.0
unemployed	112	259	34	186	142	0.0	262	0.0

Inferring Education from Jobs: From the above table it can be seen that people with management will usually have a university degree, so we can replace the 'unknown' with 'university degree'. Similarly job with 'services' education as 'high.school', job with 'housemaid' education as 'basic.4y'.

Similarly we can also infer the jobs from education where 'Education' = 'basic.4y' or 'basic.6y' or 'basic.9y' with job as 'blue-collar', if Education is 'professional.course' the job = 'technician'.

It would also make sense to replace the unknown values for job where age > 60 as 'retired'.

```python
df.loc[(df.Age>60) & (df.Job.isnull()) ,'Job'] = 'retired'
df.loc[(df.Education.isnull()) & (df.Job=='management'), 'Education'] = 'university.degree'
df.loc[(df.Education.isnull()) & (df.Job=='services'), 'Education'] = 'high.school'
df.loc[(df.Education.isnull()) & (df.Job=='housemaid'), 'Education'] = 'basic.4y'
df.loc[(df.Job.isnull()) & (df.Education=='basic.4y'), 'Job'] = 'blue-collar'
df.loc[(df.Job.isnull()) & (df.Education=='basic.6y'), 'Job'] = 'blue-collar'
df.loc[(df.Job.isnull()) & (df.Education=='basic.9y'), 'Job'] = 'blue-collar'
df.loc[(df.Job.isnull()) & (df.Education=='professional.course'), 'Job'] = 'technician'
```

Numerical Variables:

Let see the summary of data in order to understand the numerical vairables.

```
numerical_variables = ['Age', 'Campaign', 'PreviousDay', 'PrevContNum', 'EmpVarRate', 'ConsumerPriceIdx', 'ConsConfIdx', 'Euribor
df[numerical_variables].describe()
```

	Age	Campaign	PreviousDay	PrevContNum	EmpVarRate	ConsumerPriceIdx	ConsConfIdx	Euribor	Employeeno
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	2.567593	962.475454	0.172963	0.081886	93.575664	-40.502600	3.621291	5167.035911
std	10.42125	2.770014	186.910907	0.494901	1.570960	0.578840	4.628198	1.734447	72.251528
min	17.00000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	32.00000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	38.00000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	47.00000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	98.00000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000

Missing Values: From the available dataset description, missing values or NaNs are encoded as '999'. From the above screen it is clear that only PreviousDay has majority of missing values.
To deal with this variable, we will remove numerical variable PreviousDay and replace with additional categorical variables as following categories: pdays_missing (0 for contacted before and 1 for not previously contacted) , pdays_less_5, pdays_betw_5_15 and pdays_greater_15.

```
df['pdays_missing'] = 0
df['pdays_less_5'] = 0
df['pdays_betw_5_15'] = 0
df['pdays_greater_15'] = 0
df['pdays_missing'][df['PreviousDay']==999] = 1
df['pdays_less_5'][df['PreviousDay']<5] = 1
df['pdays_betw_5_15'][(df['PreviousDay']>=5) & (df['PreviousDay']<=15)] = 1
df['pdays_greater_15'][(df['PreviousDay']>15) & (df['PreviousDay'] < 999)] = 1
df_dropped_pdays = df.drop('PreviousDay', axis=1)
```