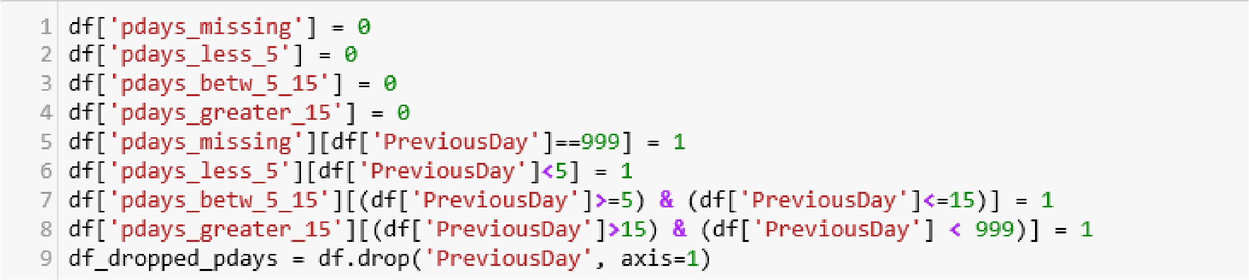
Inferential Statistics Analysis

After doing the initial exploratory data analysis , applying data wrangling techniques, we will apply some statistics concepts to further analyze data. To apply machine learning algorithms, we have to define the features and target variable. Before we apply these techniques we will check the correlation of our target variable which is whether the customer is willing to open a Term deposit or not and do some null hypothesis testing on variables as part of statistical analysis.

As explained in dataset most of the values for pdays where it is 999 are to be considered as missing values, to deal with this we will divide the pdays in different categories and have 0 or 1 value and we can drop the column after that.

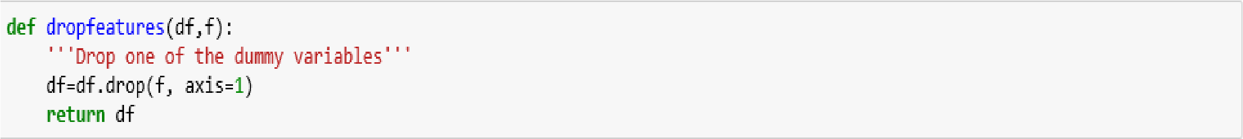


```
df['pdays_missing'] = 0
df['pdays_less_5'] = 0
df['pdays_betw_5_15'] = 0
df['pdays_greater_15'] = 0
df['pdays_missing'][df['PreviousDay']==999] = 1
df['pdays_less_5'][df['PreviousDay']<5] = 1
df['pdays_betw_5_15'][(df['PreviousDay']>=5) & (df['PreviousDay']<=15)] = 1
df['pdays_greater_15'][(df['PreviousDay']>15)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

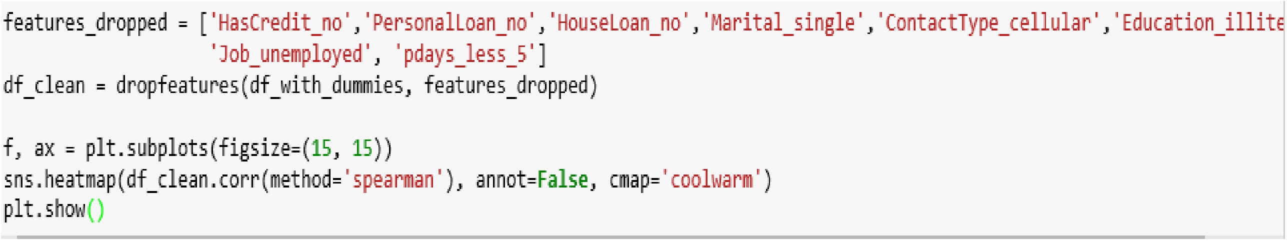
Since we have so many categories, we will create dummy variables to have a good visual correlation graph.

df_with_dummies=pd.get_dummies(df_dropped_pdays)

We will drop one of the dummy variables are for n categories we will have only n-1 variables.



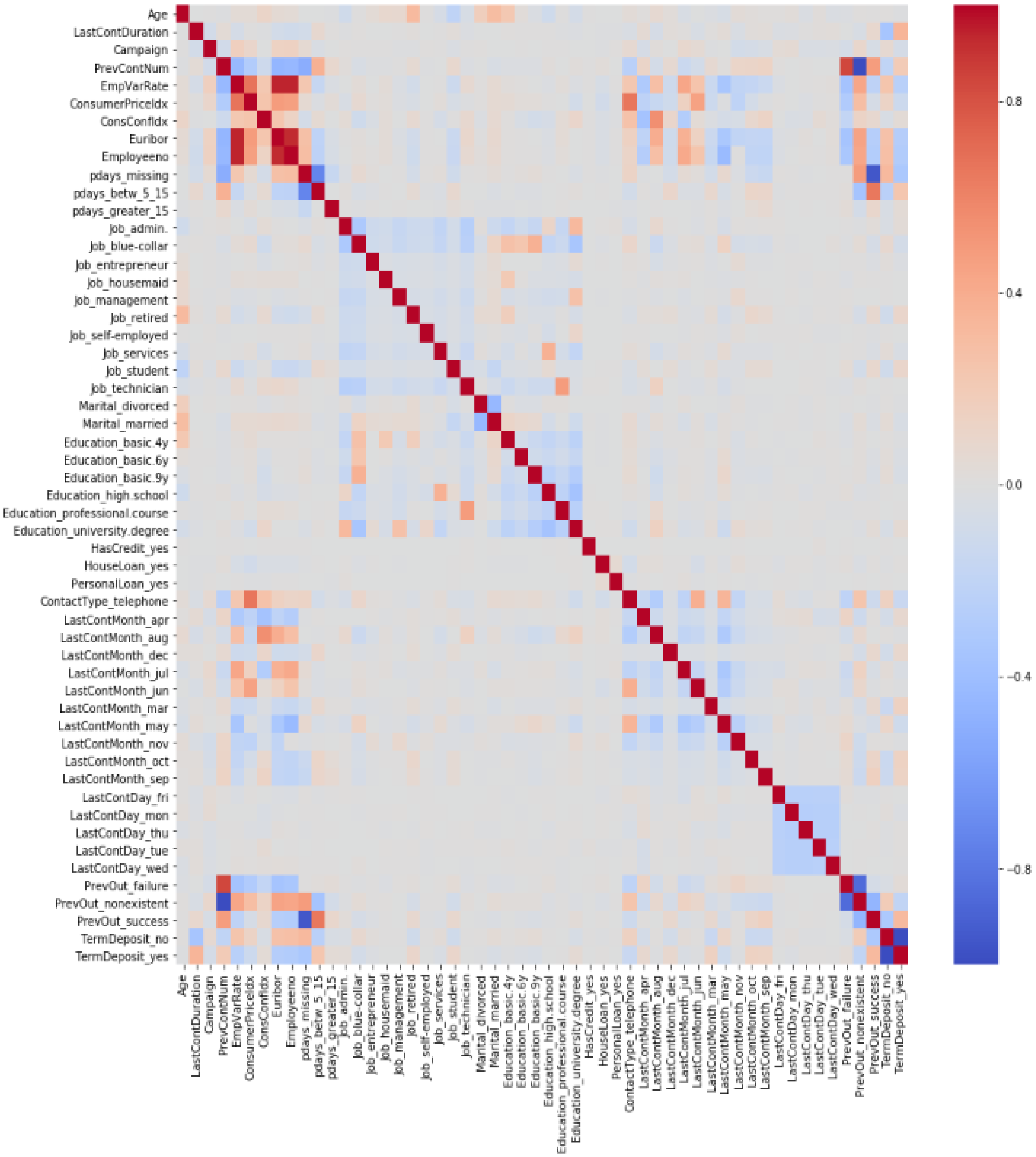
```
def dropfeatures(df,f):
    '''Drop one of the dummy variables'''
    df=df.drop(f, axis=1)
    return df
```



```
features_dropped = ['HasCredit_no','PersonalLoan_no','HouseLoan_no','Marital_single','ContactType_cellular','Education_illite
                    'Job_unemployed', 'pdays_less_5']
df_clean = dropfeatures(df_with_dummies, features_dropped)

f, ax = plt.subplots(figsize=(15, 15))
sns.heatmap(df_clean.corr(method='spearman'), annot=False, cmap='coolwarm')
plt.show()
```

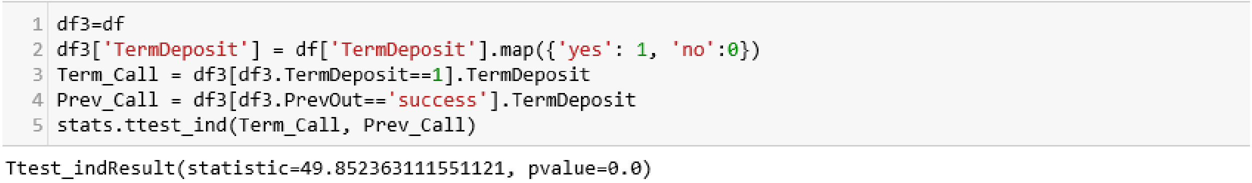
Lets plot the correlation heatmap for the df_clean to see better conclusion.



Inferences: From the above heat map we can see that 'TermDeposit_yes' (our target variable) has good correlation with **'LastContDuration', 'EmpVarRate', 'Euribor', 'Employeeno', 'pdays_missing', 'pdays_betw_5_15', 'Prevout_nonexistent' and 'Prevout_success'**. We expect to see these independent variables as significant while building the models.

Null and Alternate Hypothesis: We will perform Null hypothesis for one of the variable which is Prevout , customer who already have Term Deposit likely to open another.

Null Hypothesis: Customer who opened the previous term deposit are likely have another.
Alternate Hypothesis: There is no correlation between previous outcome success and current term deposit.



```
df3=df
df3['TermDeposit'] = df['TermDeposit'].map({'yes': 1, 'no':0})
Term_Call = df3[df3.TermDeposit==1].TermDeposit
Prev_Call = df3[df3.PrevOut=='success'].TermDeposit
stats.ttest_ind(Term_Call, Prev_Call)
```

```
Ttest_indResult(statistic=49.852363111551121, pvalue=0.0)
```

Since the pvalue is 0.0 we can reject the Null hypothesis, there might not be the complete correlation on previous customers who already have term deposit likely to open another, there can be other factors to be considered for this target variable.