Dealing with Missing Variables:

# Application:

* Variables dropped as it contains more than 50% missing data:
  + OWN\_CAR\_AGE
  + EXT\_SOURCE\_1
  + EXT\_SOURCE\_3
  + 'APARTMENTS\_AVG'
  + 'BASEMENTAREA\_AVG’
  + YEARS\_BEGINEXPLUATATION\_AVG
  + 'YEARS\_BUILD\_AVG
  + COMMONAREA\_AVG
  + 'ELEVATORS\_AVG'
  + 'ENTRANCES\_AVG'
  + 'FLOORSMIN\_AVG'
  + 'FLOORSMAX\_AVG'
  + 'LANDAREA\_AVG'
  + 'LIVINGAPARTMENTS\_AVG'
  + 'LIVINGAREA\_AVG'
  + 'NONLIVINGAPARTMENTS\_AVG'
  + 'NONLIVINGAREA\_AVG'
  + APARTMENTS\_MODE
  + BASEMENTAREA\_MODE
  + YEARS\_BEGINEXPLUATATION\_MODE
  + 'YEARS\_BUILD\_MODE'
  + COMMONAREA\_MODE
  + 'ELEVATORS\_MODE'
  + 'ENTRANCES\_MODE'
  + FLOORSMAX\_MODE
  + FLOORSMIN\_MODE'
  + LIVINGAPARTMENTS\_MODE
  + LIVINGAREA\_MODE
  + NONLIVINGAPARTMENTS\_MODE
  + 'NONLIVINGAREA\_MODE'
  + 'NONLIVINGAREA\_MODE'
  + 'APARTMENTS\_MEDI'
  + 'BASEMENTAREA\_MEDI'
  + YEARS\_BEGINEXPLUATATION\_MEDI
  + 'YEARS\_BUILD\_MEDI'
  + 'ELEVATORS\_MEDI'
  + 'ENTRANCES\_MEDI'
  + FLOORSMAX\_MEDI
  + 'FLOORSMIN\_MEDI'
  + 'LANDAREA\_MEDI'
  + 'LIVINGAPARTMENTS\_MEDI'
  + 'LIVINGAREA\_MEDI
  + NONLIVINGAPARTMENTS\_MEDI
  + NONLIVINGAREA\_MEDI
  + FONDKAPREMONT\_MODE'
  + HOUSETYPE\_MODE
  + TOTALAREA\_MODE
  + WALLSMATERIAL\_MODE'
  + 'EMERGENCYSTATE\_MODE'
* Variables dropped as it had constant data:
  + FLAG\_MOBIL
* Missing value in these variables were replaced with mean as the data are stored in a continuous distribution:
  + 'AMT\_ANNUITY', 'AMT\_GOODS\_PRICE', 'EXT\_SOURCE\_2',

'OBS\_30\_CNT\_SOCIAL\_CIRCLE', 'DEF\_30\_CNT\_SOCIAL\_CIRCLE',

'OBS\_60\_CNT\_SOCIAL\_CIRCLE', 'DEF\_60\_CNT\_SOCIAL\_CIRCLE',

'DAYS\_LAST\_PHONE\_CHANGE', 'AMT\_REQ\_CREDIT\_BUREAU\_HOUR',

'AMT\_REQ\_CREDIT\_BUREAU\_DAY', 'AMT\_REQ\_CREDIT\_BUREAU\_WEEK',

'AMT\_REQ\_CREDIT\_BUREAU\_MON', 'AMT\_REQ\_CREDIT\_BUREAU\_QRT',

'AMT\_REQ\_CREDIT\_BUREAU\_YEAR'

* Missing values replaced with most frequent observation for categorical variables:
  + NAME\_TYPE\_SUITE', 'OCCUPATION\_TYPE',

# Bureau Balance:

* No Missing variables - Following variables are removed due to the them not being significant for analysis:
  + ‘MONTHS\_BALANCE’

# Bureau:

* No Missing variables - Following variables are removed due to the them not being significant for analysis:
  + CREDIT\_CURRENCY, DAYS\_CREDIT,'CREDIT\_DAY\_OVERDUE',

'DAYS\_CREDIT\_ENDDATE', 'DAYS\_ENDDATE\_FACT', 'DAYS\_CREDIT\_UPDATE', 'AMT\_ANNUITY', ‘STATUS\_0’

* Missing value in these variables were replaced with mean as the data are stored in a continuous distribution:
  + CREDIT\_ACTIVE, AMT\_CREDIT\_SUM, AMT\_CREDIT\_SUM\_DEBT, AMT\_CREDIT\_SUM\_LIMIT, AMT\_CREDIT\_SUM\_OVERDUE, CREDIT\_TYPE\_CREDIT\_CARD, CREDIT\_TYPE\_CONSUMER\_CREDIT, AMT\_CREDIT\_MAX\_OVERDUE, CNT\_CREDIT\_PROLONG

# Installments\_payments:

* Missing values replaced with most frequent observation for categorical variables:
  + AMT\_PAYMENT\_GREATER\_EQUAL\_INSTALMENT
* Variables dropped as it contains more than 50% missing data:
  + 'DAYS\_ENTRY\_PAYMENT', 'AMT\_PAYMENT'

# Credit\_card\_balance

* Imbalanced variable (mostly contain 0) kept for now:
  + 'AMT\_DRAWINGS\_ATM\_CURRENT', 'AMT\_DRAWINGS\_CURRENT',

'AMT\_DRAWINGS\_OTHER\_CURRENT', 'AMT\_DRAWINGS\_POS\_CURRENT',

'AMT\_RECIVABLE', 'AMT\_TOTAL\_RECEIVABLE'

# POS\_CASH\_balance:

* Imbalanced variable (mostly contain 0) kept for now:
  + 'SK\_DPD', 'CNT\_INSTALMENT\_FUTURE'