```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
def import_housing_data(url):
  import pandas as pd
  df = pd.read_csv(url)
  df.drop(columns=['Id'])
  return df
  df =
import_housing_data('http://www.ishelp.info/data/housing_full.csv')
  df.head()


def unistats(df):
  import pandas as pd
  output_df =
pd.DataFrame(columns=['Count','Missing','Unique','Dtype','Mean','Mode'
,'Min','25%','Median','75%','Max','Std','Skew','Kurt'])

  for col in df:
    if pd.api.types.is_numeric_dtype(df[col]):
      output_df.loc[col] =
[df[col].count(),df[col].isnull().sum(),df[col].nunique,df[col].nuniqu
e(),df[col].dtype,pd.api.types.is_numeric_dtype(df[col]),df[col].mode(
).values[0],

df[col].mean(),df[col].mean(),df[col].min(),df[col].quantile(0.25),df[
col].median(),df[col].quantile(0.75),

df[col].max(),df[col].std(),df[col].skew(),df[col].kurt()]
    else:
      output_df.loc[col] =
[df[col].count(),df[col].isnull().sum(),df[col].nunique,df[col].nuniqu
e(),df[col].dtype,pd.api.types.is_numeric_dtype(df[col]),df[col].mode(
).values[0],
                              '-','-','-','-''-','-','-','-','-']
  return
output_df.sort_values(by=['Numeric','Unique'],ascending=False)
  #Test the Function
  import pandas as pf
  pandas.set_option('display.max_rows',100)
  pandas.set_option('display.max_columns',100)
  df = pd.read_csv('http://www.ishelp.info/data/housing_full.csv')
  unistats(df)

def anova(df, feature, label):
  import pandas as pd
  import numpy as np
  from scipy import stats
```

```
  groups = df[feature].unique()
  df_grouped = df.groupby(feature)
  group_labels = []
  for g in groups:
   g_list = df_grouped.get_group(g)
   group_labels.append(g_list[label])

  return stats.f_oneway(*group_labels)

# Bivariate: Numeric to numeric: Correlation
# Bivariate: Numeric to categorical: one-way ANOVA (3+ groups) or t-
test (2 groups)
# Bivariate: categorical to categorical: Chi-square

def bivstats(df, label):
  from scipy import stats
  import pandas as pd
  import numpy as np

  # Create an empty DataFrame to store output
  output_df = pd.DataFrame(columns=['stat', '+/-', 'Effect size', 'p-
value'])

  for col in df:
    if not col == label:
      if df[col].isnull().sum() == 0:
        if pd.api.types.is_numeric_dtype(df[col]):
          r, p = stats.pearsonr(df[label], df[col])
          output_df.loc[col] = ['r', np.sign(r), abs(round(r, 3)),
round(p, 6)]
        else:
          F, p = anova(df[[col, label]], col, label)
          output_df.loc[col] = ['F', '', round(F, 3), round(p, 6)]
      else:
        output_df.loc[col] = [np.nan, np.nan, np.nan, 'nulls']

  return output_df.sort_values(by=['Effect size', 'stat'],
ascending=[False, False])

import pandas as pd
pd.options.display.float_format = '{:.5f}'.format
df = pd.read_csv('http://www.ishelp.info/data/housing_full.csv')
bivstats(df, 'SalePrice')
```

```
            stat  +/-  Effect size p-value
ExterQual      F          443.33500 0.00000
KitchenQual    F          407.80600 0.00000
Foundation     F          100.25400 0.00000
CentralAir     F           98.30500 0.00000
```

```
HeatingQC      F          88.39400 0.00000
...          ...  ...          ...     ...
GarageQual   NaN  NaN          NaN   nulls
GarageCond   NaN  NaN          NaN   nulls
PoolQC       NaN  NaN          NaN   nulls
Fence        NaN  NaN          NaN   nulls
MiscFeature  NaN  NaN          NaN   nulls

[80 rows x 4 columns]
```

```
def import_housing_data(url):
  df = pd.read_csv(url)
  df.drop(columns=['Id'], inplace=True)
  df.dropna(axis=1, inplace=True)

  for col in df:
    if col[0].isdigit():
      nums = ['zero', 'one', 'two', 'three', 'four', 'five', 'six',
'seven', 'eight', 'nine']
      df. rename(columns={col:nums[int(col[0])] + '_' + col},
inplace=True)

  return df

import sys
sys.path.append('/content/drive/My Drive/ColabNotebooks/')
import pandas as pd
pd.set_option('display.max_rows', 1000)
pd.set_option('display.max_columns', 100)
pd.options.display.float_format = '{:.8f}'.format

df = import_housing_data('http://ishelp.info/data/housing_full.csv')
df.head()
```

```
   MSSubClass MSZoning  LotArea Street LotShape LandContour Utilities
\
0          60       RL     8450   Pave      Reg         Lvl    AllPub

1          20       RL     9600   Pave      Reg         Lvl    AllPub

2          60       RL    11250   Pave      IR1         Lvl    AllPub

3          70       RL     9550   Pave      IR1         Lvl    AllPub

4          60       RL    14260   Pave      IR1         Lvl    AllPub


  LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
HouseStyle  \
```

```
0    Inside       Gtl      CollgCr       Norm       Norm     1Fam
2Story
1       FR2       Gtl      Veenker      Feedr       Norm     1Fam
1Story
2    Inside       Gtl      CollgCr       Norm       Norm     1Fam
2Story
3    Corner       Gtl      Crawfor       Norm       Norm     1Fam
2Story
4       FR2       Gtl      NoRidge       Norm       Norm     1Fam
2Story

   OverallQual  OverallCond  YearBuilt  YearRemodAdd RoofStyle
RoofMatl  \
0            7            5       2003          2003     Gable
CompShg
1            6            8       1976          1976     Gable
CompShg
2            7            5       2001          2002     Gable
CompShg
3            7            5       1915          1970     Gable
CompShg
4            8            5       2000          2000     Gable
CompShg

  Exterior1st Exterior2nd ExterQual ExterCond Foundation
BsmtFinSF1  \
0     VinylSd     VinylSd        Gd        TA      PConc         706

1     MetalSd     MetalSd        TA        TA     CBlock         978

2     VinylSd     VinylSd        Gd        TA      PConc         486

3     Wd Sdng     Wd Shng        TA        TA     BrkTil         216

4     VinylSd     VinylSd        Gd        TA      PConc         655


   BsmtFinSF2  BsmtUnfSF  TotalBsmtSF Heating HeatingQC CentralAir  \
0           0        150          856    GasA        Ex          Y
1           0        284         1262    GasA        Ex          Y
2           0        434          920    GasA        Ex          Y
3           0        540          756    GasA        Gd          Y
4           0        490         1145    GasA        Ex          Y

   one_1stFlrSF  two_2ndFlrSF  LowQualFinSF  TotalSF  BsmtFullBath  \
0           856           854             0     1710             1
1          1262             0             0     1262             0
2           920           866             0     1786             1
3           961           756             0     1717             1
```

```
4          1145          1053             0     2198             1

   BsmtHalfBath  FullBath  HalfBath  BedroomAbvGr  KitchenAbvGr
KitchenQual  \
0             0         2         1             3             1
Gd
1             1         2         0             3             1
TA
2             0         2         1             3             1
Gd
3             0         1         0             3             1
Gd
4             0         2         1             4             1
Gd

   TotRmsAbvGrd Functional  Fireplaces  GarageCars  GarageArea
PavedDrive  \
0             8        Typ           0           2         548
Y
1             6        Typ           1           2         460
Y
2             6        Typ           1           2         608
Y
3             7        Typ           1           3         642
Y
4             9        Typ           1           3         836
Y

   WoodDeckSF  OpenPorchSF  EnclosedPorch  three_3SsnPorch
ScreenPorch  \
0           0           61              0                0
0
1         298            0              0                0
0
2           0           42              0                0
0
3           0           35            272                0
0
4         192           84              0                0
0

   PoolArea  MiscVal  MoSold  YrSold SaleType SaleCondition  SalePrice

0         0        0       2    2008       WD        Normal     208500

1         0        0       5    2007       WD        Normal     181500

2         0        0       9    2008       WD        Normal     223500
```

```
3          0          0        2    2006        WD        Abnorml      140000

4          0          0       12    2008        WD         Normal      250000
```

```
def import_housing_data(df, label):
  import numpy as np
  import pandas as pd
  import statsmodels.api as sm
  from sklearn import preprocessing

  label = 'SalePrice'

  for col in df:
     if not pd.api.types.is_numeric_dtype(df[col]):
      df = df.join(pd.get_dummies(df[col], prefix=col,
drop_first=False))
      df = df.select_dtypes(np.number)

  d_minmax =
pd.DataFrame(preprocessing.MinMaxScaler().fit_transform(df),
columns=df.columns)
  y = df_minmax[label]
  x = df_minmax.drop(columns=[label, 'Utilities AllPub',
'Exteriorist_BrkComm']).assign(const=1)
  results = sm.OLS(y, x).fit()
  results.summary()
```