# -\*- coding: utf-8 -\*-

"""

Created on Mon Nov 22 19:20:55 2021

@author: Hydraconix

"""

#Standard Data Science Imports

import numpy as np

import pandas as pd

from pandas import Series, DataFrame

# Visualization libraries

import seaborn as sns

import matplotlib.pyplot as plt

%matplotlib inline

# Statistics packages

import pylab

from statsmodels.formula.api import ols

import statistics

from scipy import stats

# Scikit-learn

import sklearn

from sklearn import preprocessing

from sklearn.preprocessing import LabelEncoder

from sklearn.linear\_model import LinearRegression

from sklearn import metrics

# Load data set into Pandas Dataframe

churn\_df = pd.read\_csv(r'C:\Users\Hydraconix\Desktop\churn\_clean.csv')

# Checking for Null Values

churn\_df.isna().sum()

#Summary Statistics

churn\_df.Age.describe()

churn\_df.Children.describe()

churn\_df.Income.describe()

churn\_df.Outage\_sec\_perweek.describe()

churn\_df.Yearly\_equip\_failure.describe()

churn\_df.Tenure.describe()

churn\_df.MonthlyCharge.describe()

churn\_df.Bandwidth\_GB\_Year.describe()

# Rename Last 8 Survey Columns for better description of variables

churn\_df.rename(columns = {'Item1' : 'TimelyResponse',

'Item2' : 'Fixes' ,

'Item3' : 'Replacements' ,

'Item4' : 'Reliability' ,

'Item5' : 'Options' ,

'Item6' : 'Respectfulness' ,

'Item7' : 'Courteous' ,

'Item8' : 'Listening'},

inplace=True)

# Remove less meaningful demographic variables from statistics description

churn\_df = churn\_df.drop(columns=['CaseOrder' ,

'Customer\_id' ,

'Interaction' ,

'UID' ,

'City' ,

'State' ,

'County' ,

'Zip' ,

'Lat' ,

'Lng' ,

'Population' ,

'Area' ,

'TimeZone' ,

'Job' ,

'Marital' ,

'PaymentMethod'])

# Converting binary categorical variables to numeric variables

churn\_df['DummyChurn'] = [1 if v == 'Yes' else 0 for v in churn\_df['Churn']]

churn\_df['DummyTechie'] = [1 if v == 'Yes' else 0 for v in churn\_df['Techie']]

churn\_df['DummyPort\_modem'] = [1 if v == 'Yes' else 0 for v in churn\_df['Port\_modem']]

churn\_df['DummyTablet'] = [1 if v == 'Yes' else 0 for v in churn\_df['Tablet']]

churn\_df['DummyPhone'] = [1 if v == 'Yes' else 0 for v in churn\_df['Phone']]

churn\_df['DummyMultiple'] = [1 if v == 'Yes' else 0 for v in churn\_df['Multiple']]

churn\_df['DummyOnlineSecurity'] = [1 if v == 'Yes' else 0 for v in churn\_df['OnlineSecurity']]

churn\_df['DummyOnlineBackup'] = [1 if v == 'Yes' else 0 for v in churn\_df['OnlineBackup']]

churn\_df['DummyDeviceProtection'] = [1 if v == 'Yes' else 0 for v in churn\_df['DeviceProtection']]

churn\_df['DummyTechSupport'] = [1 if v == 'Yes' else 0 for v in churn\_df['TechSupport']]

churn\_df['DummyStreamingTV'] = [1 if v == 'Yes' else 0 for v in churn\_df['StreamingTV']]

churn\_df['DummyStreamingMovies'] = [1 if v == 'Yes' else 0 for v in churn\_df['StreamingMovies']]

churn\_df['DummyPaperlessBilling'] = [1 if v == 'Yes' else 0 for v in churn\_df['PaperlessBilling']]

# Converting ordinal categorical data into numeric variables

churn\_df['DummyInternetService'] = churn\_df.InternetService.map({'None' : 0, 'DSL' : 1, 'Fiber Optic' : 2})

churn\_df['DummyContract'] = churn\_df.Contract.map({'Month-to-month' : 0, 'One year' : 1, 'Two Year' : 2})

churn\_df['DummyGender'] = churn\_df.Gender.map({'Nonbinary' : 0, 'Male' : 1, 'Female' : 2})

# Drop original categorical variables from dataframe

churn\_df = churn\_df.drop(columns=['Gender' ,

'Churn' ,

'Techie' ,

'Contract' ,

'Port\_modem' ,

'Tablet' ,

'InternetService' ,

'Phone' ,

'Multiple' ,

'OnlineSecurity' ,

'OnlineBackup',

'DeviceProtection' ,

'TechSupport' ,

'StreamingTV',

'StreamingMovies',

'PaperlessBilling'])

#Create histograms of continuous variables

churn\_df[['Children',

'Age' ,

'Income' ,

'Outage\_sec\_perweek' ,

'Email' ,

'Contacts' ,

'Yearly\_equip\_failure' ,

'Tenure' , 'MonthlyCharge' ,

'Bandwidth\_GB\_Year' ,

‘DummyGender’ ,

‘DummyInternetService’ ,

‘DummyContract’]].hist()

plt.savefig('churn\_pyplot.jpg')

plt.tight\_layout()

# Create Seaborn boxplots for continuous variables

sns.boxplot('Tenure' , data = churn\_df)

plt.show()

sns.boxplot('MonthlyCharge' , data = churn\_df)

plt.show()

sns.boxplot('Bandwidth\_GB\_Year' , data = churn\_df)

plt.show()

# Run Scatterplots to show direct or inverse relationships between the target & independent variables

sns.scatterplot(x=churn\_df['Children'], y=churn\_df['Bandwidth\_GB\_Year'],

color='red')

plt.show()

sns.scatterplot(x=churn\_df['Age'], y=churn\_df['Bandwidth\_GB\_Year'],

color='red')

plt.show()

sns.scatterplot(x=churn\_df['Income'], y=churn\_df['Bandwidth\_GB\_Year'],

color='red')

plt.show()

sns.scatterplot(x=churn\_df['Outage\_sec\_perweek'], y=churn\_df['Bandwidth\_GB\_Year'],

color='red')

plt.show()

sns.scatterplot(x=churn\_df['Email'], y=churn\_df['Bandwidth\_GB\_Year'],

color='red')

plt.show()

sns.scatterplot(x=churn\_df['Contacts'], y=churn\_df['Bandwidth\_GB\_Year'],

color='red')

plt.show()

sns.scatterplot(x=churn\_df['Yearly\_equip\_failure'], y=churn\_df['Bandwidth\_GB\_Year'],

color='red')

plt.show()

sns.scatterplot(x=churn\_df['Tenure'], y=churn\_df['Bandwidth\_GB\_Year'],

color='red')

plt.show()

sns.scatterplot(x=churn\_df['MonthlyCharge'], y=churn\_df['Bandwidth\_GB\_Year'],

color='red')

plt.show()

sns.scatterplot(x=churn\_df['TimelyResponse'], y=churn\_df['Bandwidth\_GB\_Year'],

color='red')

plt.show()

sns.scatterplot(x=churn\_df['Fixes'], y=churn\_df['Bandwidth\_GB\_Year'],

color='red')

plt.show()

sns.scatterplot(x=churn\_df['DummyTechie'], y=churn\_df['Bandwidth\_GB\_Year'],

color='red')

plt.show()

sns.scatterplot(x=churn\_df['DummyGender'], y=churn\_df['Bandwidth\_GB\_Year'],

color='red')

plt.show()

sns.scatterplot(x=churn\_df['DummyContract'], y=churn\_df['Bandwidth\_GB\_Year'],

color='red')

plt.show()

sns.scatterplot(x=churn\_df['DummyInternetService'], y=churn\_df['Bandwidth\_GB\_Year'],

color='red')

plt.show()

# Provide a copy of the prepared data set

churn\_df.to\_csv('churn\_prepared.csv')

df = churn\_df.columns

print(df)

# Develop the initial estimated regression equation that could be used to predict the Bandwidth\_GB\_Year, given the only continuous variables

lm\_bandwidth = ols("Bandwidth\_GB\_Year ~ Children + Age + Income + Outage\_sec\_perweek + Email +Contacts +Yearly\_equip\_failure +Tenure + MonthlyCharge + TimelyResponse + Fixes + Replacements + Reliability + Options + Respectfulness + Courteous + Listening", data=churn\_df).fit()

print(lm\_bandwidth.params)

print(lm\_bandwidth.summary())

# Model including all dummy variables

lm\_bandwidth = ols("Bandwidth\_GB\_Year ~ Children + Age + Income + Outage\_sec\_perweek + Email +Contacts +Yearly\_equip\_failure +Tenure + MonthlyCharge + TimelyResponse + Fixes + Replacements + Reliability + Options + Respectfulness + Courteous + Listening + DummyChurn + DummyTechie + DummyPort\_modem + DummyTablet + DummyPhone + DummyMultiple + DummyOnlineSecurity + DummyOnlineBackup + DummyDeviceProtection + DummyTechSupport + DummyStreamingTV + DummyStreamingMovies + DummyPaperlessBilling + DummyInternetService + DummyContract + DummyGender", data=churn\_df).fit()

print(lm\_bandwidth.params)

print(lm\_bandwidth.summary())

# Create dataframe for heatmap bivariate analysis of correlation

churn\_bivariate = churn\_df[['Bandwidth\_GB\_Year',

'Children',

'Age',

'Income',

'Outage\_sec\_perweek',

'Yearly\_equip\_failure',

'DummyTechie',

'DummyContract',

'DummyPort\_modem',

'DummyTablet',

'DummyInternetService',

'DummyPhone',

'DummyMultiple',

'DummyOnlineSecurity',

'DummyOnlineBackup',

'DummyDeviceProtection',

'DummyTechSupport',

'DummyStreamingTV',

'DummyPaperlessBilling',

'DummyGender' ,

'DummyStreamingMovies' ,

'Email',

'Contacts',

'Tenure',

'MonthlyCharge',

'TimelyResponse',

'Fixes',

'Replacements',

'Reliability',

'Options',

'Respectfulness',

'Courteous',

'Listening']]

# Run Seaborn heatmap

sns.heatmap(churn\_bivariate.corr(), annot=False)

plt.show()

# Run reduced OLS multiple regression

lm\_bandwidth\_reduced = ols("Bandwidth\_GB\_Year ~ Children + Tenure + Fixes + Replacements", data=churn\_df).fit()

print(lm\_bandwidth.params)

print(lm\_bandwidth.summary())

churn\_df = pd.read\_csv('churn\_prepared.csv')

residuals = churn\_df['Bandwidth\_GB\_Year'] = lm\_bandwidth\_reduced.predict(churn\_df[['Children', 'Tenure', 'Fixes','Replacements']])

sns.scatterplot(x=churn\_df['Tenure'],y=residuals,color='red')

plt.show()