Predicting Customer Tenure With Multipe Linear Regression

```
import numpy as np
import pandas as pd
import matplotlib
%matplotlib inline
from matplotlib import pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_regression import LinearRegression
import sklearn.metrics
```

Data Aquisition

```
# Import the data set
df = pd.read_csv('/home/jovyan/work/datasets/Telco-Customer-Churn.csv')
df.shape, df.dtypes, df.head(5), df.tail(5)
```

```
# Delete a few duplicated rows and the tenure observations of zero(see notebook)
df = df[df.TotalCharges != 0.0]
df.shape, df.dtypes, df.head(5), df.tail(5)
```

Data Exploration

```
# Produce statistics for continuous variables
df[['MonthlyCharges']].describe()
print('The mean of MonthlyCharges is', df.MonthlyCharges.mean())
print('The standard deviation of MonthlyCharges is', np.std(df.MonthlyCharges))
print('The median of MonthlyCharges is', df.MonthlyCharges.median())
print('The range of MonthlyCharges is', df.MonthlyCharges.ptp())
print('The mode of MonthlyCharges is', df.MonthlyCharges.mode())
print('The kurtosis of MonthlyCharges is', df.MonthlyCharges.kurtosis())
print('The skewness of MonthlyCharges is', df.MonthlyCharges.skew())
```

```
MonthlyCharges
count  6530.000000
mean     64.761692
std      30.090047
min      18.250000
25%      35.587500
50%      70.350000
75%      89.862500
max     118.750000
dtype: float64
The mean of MonthlyCharges is 64.7617915309446
The standard deviation of MonthlyCharges is 30.08405319031311
The median of MonthlyCharges is 70.35
The range of MonthlyCharges is 100.5
The mode of MonthlyCharges is 0    20.05
dtype: float64
The kurtosis of MonthlyCharges is -0.9135311160917835
The skewness of MonthlyCharges is 0.22032501309231114
```

```
# Produce statistics for continuous variables
df[['TotalCharges']].describe()
print('The mean of TotalCharges is', df.TotalCharges.mean())
print('The standard deviation of TotalCharges is', np.std(df.TotalCharges))
print('The median of TotalCharges is', df.TotalCharges.median())
print('The range of TotalCharges is', df.TotalCharges.ptp())
print('The mode of TotalCharges is', df.TotalCharges.mode())
print('The kurtosis of TotalCharges is', df.TotalCharges.kurtosis())
print('The skewness of TotalCharges is', df.TotalCharges.skew())
```

```
TotalCharges
count  6530.000000
mean   2283.300441
std    2266.771362
min      18.800000
25%     401.450000
50%    1397.475000
75%    3794.737500
max    8684.800000
dtype: float64
The mean of TotalCharges is 2283.3004408418657
The standard deviation of TotalCharges is 2266.7713616121114
The median of TotalCharges is 1397.475
The range of TotalCharges is 8665.9999999999998
The mode of TotalCharges is 0    1889.5
dtype: float64
The kurtosis of TotalCharges is -0.9881365966233413
The skewness of TotalCharges is 0.9615650114179187
```

```
# Produce statistics for continuous variables
df[['tenure']].describe()
print('The mean of tenure is', df.tenure.mean())
print('The standard deviation of tenure is', np.std(df.tenure))
print('The median of tenure is', df.tenure.median())
print('The range of tenure is', df.tenure.ptp())
print('The mode of tenure is', df.tenure.mode())
print('The kurtosis of tenure is', df.tenure.kurtosis())
print('The skewness of tenure is', df.tenure.skew())
```

```
tenure
count  6530.000000
mean     32.421516
std      24.545260
min       1.000000
25%       9.000000
50%      29.000000
75%      55.000000
max      72.000000
dtype: float64
The mean of tenure is 32.421516079632465
The standard deviation of tenure is 24.537594319457013
The median of tenure is 29.0
The range of tenure is 71.0
The mode of tenure is 0    1
dtype: float64
The kurtosis of tenure is -1.3879965202368363
The skewness of tenure is 0.2395398501194592
```

```
# Produce statistics for continuous variables
df[['SeniorCitizen']].describe()
print('The mean of SeniorCitizen is', df.SeniorCitizen.mean())
print('The standard deviation of SeniorCitizen is', np.std(df.SeniorCitizen))
print('The median of SeniorCitizen is', df.SeniorCitizen.median())
print('The range of SeniorCitizen is', df.SeniorCitizen.ptp())
print('The mode of SeniorCitizen is', df.SeniorCitizen.mode())
print('The kurtosis of SeniorCitizen is', df.SeniorCitizen.kurtosis())
print('The skewness of SeniorCitizen is', df.SeniorCitizen.skew())
```

```
SeniorCitizen
count  6530.000000
mean     0.161562
std      0.367909
min      0.000000
25%      0.000000
50%      0.000000
75%      0.000000
max      1.000000
dtype: float64
The mean of SeniorCitizen is 0.1615620214395099
The standard deviation of SeniorCitizen is 0.3679088274999411
The median of SeniorCitizen is 0.0
The range of SeniorCitizen is 1.0
The mode of SeniorCitizen is 0    0
dtype: int64
The kurtosis of SeniorCitizen is 1.3548777160842372
The skewness of SeniorCitizen is 1.8341759641496457
```

```
# Produce statistics for continuous variables
df[['MonthlyCharges']].describe()
print('The mean of MonthlyCharges is', df.MonthlyCharges.mean())
print('The standard deviation of MonthlyCharges is', np.std(df.MonthlyCharges))
print('The median of MonthlyCharges is', df.MonthlyCharges.median())
print('The range of MonthlyCharges is', df.MonthlyCharges.ptp())
print('The mode of MonthlyCharges is', df.MonthlyCharges.mode())
print('The kurtosis of MonthlyCharges is', df.MonthlyCharges.kurtosis())
print('The skewness of MonthlyCharges is', df.MonthlyCharges.skew())
```

```
MonthlyCharges
count  6530.000000
mean     64.761692
std      30.090047
min      18.250000
25%      35.587500
50%      70.350000
75%      89.862500
max     118.750000
dtype: float64
The mean of MonthlyCharges is 64.7617915309446
The standard deviation of MonthlyCharges is 30.08405319031311
The median of MonthlyCharges is 70.35
The range of MonthlyCharges is 100.5
The mode of MonthlyCharges is 0    20.05
dtype: float64
The kurtosis of MonthlyCharges is -0.9135311160917835
The skewness of MonthlyCharges is 0.22032501309231114
```

```
# Produce statistics for continuous variables
df[['MonthlyPrice_Beyond_Mean']].describe()
print('The mean of MonthlyPrice_Beyond_Mean is', df.MonthlyPrice_Beyond_Mean.mean())
print('The standard deviation of MonthlyPrice_Beyond_Mean is', np.std(df.MonthlyPrice_Beyond_Mean))
print('The median of MonthlyPrice_Beyond_Mean is', df.MonthlyPrice_Beyond_Mean.median())
print('The range of MonthlyPrice_Beyond_Mean is', df.MonthlyPrice_Beyond_Mean.ptp())
print('The mode of MonthlyPrice_Beyond_Mean is', df.MonthlyPrice_Beyond_Mean.mode())
print('The kurtosis of MonthlyPrice_Beyond_Mean is', df.MonthlyPrice_Beyond_Mean.kurtosis())
print('The skewness of MonthlyPrice_Beyond_Mean is', df.MonthlyPrice_Beyond_Mean.skew())
```

```
MonthlyPrice_Beyond_Mean
count  6530.000000
mean     -0.000000
std      30.090047
min     -46.511692
25%     -29.174192
50%      5.588308
75%      25.100808
max      53.988308
dtype: float64
The mean of MonthlyPrice_Beyond_Mean is -1.8189894035458565e-12
The standard deviation of MonthlyPrice_Beyond_Mean is 30.08405319031311
The median of MonthlyPrice_Beyond_Mean is 5.588308
The range of MonthlyPrice_Beyond_Mean is 100.5
The mode of MonthlyPrice_Beyond_Mean is 0    -44.711692
dtype: float64
The kurtosis of MonthlyPrice_Beyond_Mean is -0.9135311160917835
The skewness of MonthlyPrice_Beyond_Mean
```

```
Figure size 1000x600 with 1 Axes
```

```
# Visualize Attend_Class distribution
# Create a count plot
plt.figure(figure_size=(10, 6))
sns.countplot(x='Attend_Class', data=df)
plt.title('Attend_Class Distribution')
plt.xlabel('Attend_Class')
plt.ylabel('Count')
plt.show()

# Visualize Age distribution
# Create a histogram
plt.figure(figure_size=(10, 6))
sns.histplot(df['Age'], bins=20, kde=True)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()

# Visualize Marital_status distribution
# Create a count plot
plt.figure(figure_size=(10, 6))
sns.countplot(x='Marital_status', data=df)
plt.title('Marital_status Distribution')
plt.xlabel('Marital_status')
plt.ylabel('Count')
plt.show()

# Visualize Gender distribution
# Create a count plot
plt.figure(figure_size=(10, 6))
sns.countplot(x='Gender', data=df)
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()

# Visualize Attend_Class distribution
# Create a count plot
plt.figure(figure_size=(10, 6))
sns.countplot(x='Attend_Class', data=df)
plt.title('Attend_Class Distribution')
plt.xlabel('Attend_Class')
plt.ylabel('Count')
plt.show()

# Visualize Age distribution
# Create a histogram
plt.figure(figure_size=(10, 6))
sns.histplot(df['Age'], bins=20, kde=True)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()

# Visualize Marital_status distribution
# Create a count plot
plt.figure(figure_size=(10, 6))
sns.countplot(x='Marital_status', data=df)
plt.title('Marital_status Distribution')
plt.xlabel('Marital_status')
plt.ylabel('Count')
plt.show()

# Visualize Gender distribution
# Create a count plot
plt.figure(figure_size=(10, 6))
sns.countplot(x='Gender', data=df)
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()

# Visualize Attend_Class distribution
# Create a count plot
plt.figure(figure_size=(10, 6))
sns.countplot(x='Attend_Class', data=df)
plt.title('Attend_Class Distribution')
plt.xlabel('Attend_Class')
plt.ylabel('Count')
plt.show
```

```
Figure size 1000x600 with 1 Axes
```

```
# Create a pairplot for numerical variables
sns.pairplot(df[['Age', 'Attend_Class']])
plt.show()

# Create a heatmap of correlations between numerical variables
correlation_matrix = df[['Age', 'Attend_Class']].corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.show()
```

Build the Model and Evaluate the Results

```
In [100]: #Develop an initial model for regression that utilizes all predictors initially identified in C2
          import statsmodels.api as sm
          X = df[['Tenure', 'Zip', 'Population', 'Children', 'Age', 'Income', 'Outage_sec_perweek', 'Email', 'Contacts', 'Yearly_equip_failure', 'MonthlyCharge', 'Area_
          model1.summary()
```

```
Out[100]:
                            OLS Regression Results
Dep. Variable:                 Tenure   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                  -0.000
Method:                 Least Squares   F-statistic:                     0.8603
Date:                Fri, 20 Aug 2021   Prob (F-statistic):              0.716
Time:                        19:10:08   Log-Likelihood:                -14181.
No. Observations:                9900   AIC:                             2.839e+04
Df Residuals:                    9879   BIC:                             2.854e+04
Df Model:                          20
Covariance Type:            nonrobust
                 coef    std err          t      P>|t|    [0.025    0.975]
------------------------------------------------------------------------------
Intercept     1.384e-16    0.012    1.09e-14      1.000    -0.025     0.025
Zip           -0.0034    0.015    -0.226      0.821    -0.033     0.026
Population    -0.0037    0.015    -0.248      0.804    -0.033     0.026
Children      -0.0047    0.015    -0.311      0.756    -0.034     0.025
Age           0.0170    0.015     1.130      0.259    -0.012     0.047
Income        0.0010    0.015     0.066      0.947    -0.028     0.031
Outage_sec_perweek    0.0016    0.015     0.108      0.914    -0.028     0.031
Email         -0.0041    0.015    -0.271      0.787    -0.034     0.025
Contacts      0.0004    0.015     0.028      0.977    -0.027     0.032
Yearly_equip_failure    0.0016    0.015     0.108      0.914    -0.028     0.031
MonthlyCharge    -0.0042    0.015    -0.281      0.779    -0.033     0.026
Area_Suburban    -0.0144    0.015    -0.946      0.344    -0.043     0.015
Area_Urban    -0.0186    0.015    -1.227      0.220    -0.048     0.011
Marital_Married    0.0142    0.015     0.931      0.351    -0.015     0.044
Marital_Married_Married    0.0062    0.015     0.407      0.684    -0.024     0.036
Marital_Separated    0.0362    0.015     2.387      0.017    0.006     0.066
Marital_Widowed    0.0095    0.015     0.627      0.530    -0.019     0.039
Gender_Male    -0.0104    0.015    -0.684      0.494    -0.040     0.019
Gender_Nonbinary    -0.0129    0.015    -0.854      0.393    -0.042     0.016
Internet Service_Fiber_Optic    -0.0162    0.015    -0.861      0.389    -0.053     0.021
Internet Service_None    -0.0158    0.015    -1.369      0.169    -0.038     0.006
Omnibus:       4043.539   Durbin-Watson:                   0.180
Prob(Omnibus):    0.000   Jarque-Bera (JB):             1298.357
Skew:             0.008   Prob(JB):                     1.24e-282
Kurtosis:         1.241   Cond. No.                         2.53
```

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```
In [101]: '''This initial model does little to explain the response variable due to the very low adjusted R-squared value
          Further exploration of the model will be needed to see if any predictors are significantly contributing
          Further exploration of the original data set is required for this process and a correlation matrix to assess
          additional predictors will be needed.
          Multicollinearity issues will be addressed subsequently added to this code for more efficient execution of repeated
          analysis'''
```

```
In [102]: #generate expanded correlation matrix to determine which items may have a stronger relationship with
          #tenure
          print(clean_churn.corr(),sep='\')
```

```
Customer_ID                 Lat      Lng  \\
Customer_ID       1.000000e-16 0.003750e-16 -0.014346 -0.001612
Zip               0.003007 1.000000e-00 -0.842580 -0.901788
Lat               -0.014346 -0.842580e-01  1.000000  0.000000
Lng               -0.001612 -0.901788e-01  0.000000  1.000000
Population        -0.007494 -0.913606e-02 -0.220395 -0.044294
Children          -0.000000 -0.000000e-00  0.000000  0.000000
Age               0.016028 -0.006262e-02  0.042671  0.012117
Income            0.001352 -0.948524e-02  0.006924 -0.002997
Outage_sec_perweek    0.000000 -0.000000e-00  0.000000  0.000000
Email             -0.011274 -0.006000e-02  0.000000  0.000000
Contacts          0.000000 -0.000000e-00  0.000000  0.000000
Yearly_equip_failure    0.005487 1.166692e-02 -0.005077 -0.006683
MonthlyCharge     0.000000 -0.000000e-00  0.000000  0.000000
Bandwidth_GB_Year    0.001967 -0.724780e-02  0.014845  0.011765
Bandwidth_GB_Year    0.825561 -2.527106e-01 0.016945  0.003446
Item1             0.000000 -0.000000e-00  0.000000  0.000000
Item2             0.000000 -0.000000e-00  0.000000  0.000000
Item3             0.002465 -0.968476e-03 0.003421  0.003866
Item4             0.000000 -0.000000e-00  0.000000  0.000000
Item5             0.014429 -0.002609e-02 0.000000  0.000000
Item6             0.000000 -0.000000e-00  0.000000  0.000000
Item7             0.002995 -5.537567e-03 0.000897  0.003401
Item8             0.014840 -2.037761e-02 0.000000  0.000000

                  Children       Age    Income  Outage_sec_perweek  \\
Customer_ID       0.000000 0.016028 0.001352            0.000000
Zip               -0.871705 0.006262 -0.948524            0.000000
Lat               0.000000 0.042671 0.006924            0.000000
Lng               0.000000 0.012117 -0.002997            0.000000
Population        -0.003670 0.010506 -0.006628            0.000000
Children          1.000000 0.000000 0.000000            0.000000
Age               0.000000 1.000000 0.000000            0.000000
Income            0.000000 0.000000 1.000000            0.000000
Outage_sec_perweek    0.000000 0.000000 0.000000            1.000000
Email             0.000000 0.000000 0.000000            0.000000
Contacts          0.000000 0.000000 0.000000            0.000000
Yearly_equip_failure    0.000000 0.000000 0.000000            0.000000
MonthlyCharge     0.000000 0.000000 0.000000            0.000000
Bandwidth_GB_Year    0.000000 0.014754 0.003476            0.000000
Bandwidth_GB_Year    0.013478 -0.000572 -0.004009            -0.021228
Item1             0.000000 0.000000 0.000000            0.000000
Item2             0.000000 0.000000 0.000000            0.000000
Item3             0.000000 0.000000 0.000000            0.000000
Item4             0.000000 -0.006305 0.011111            -0.004134
Item5             0.000000 0.000000 0.000000            0.000000
Item6             0.000000 0.000000 0.000000            0.000000
Item7             -0.000126 0.013163 0.000760            0.008126

                  Email  Contacts  Yearly_equip_failure   Tenure  \\
Customer_ID       0.011274  0.000000  0.005487  0.002000
Zip               0.000000 -0.000000 1.166692e-02 0.825561
Lat               0.000000 0.000000 -0.005077 -0.007543
Lng               0.000000 0.000000 -0.006683 -0.003446
Population        0.017842 0.000000 0.000000 -0.000000
Children          0.000000 0.000000 0.000000 0.000000
Age               0.000000 0.000000 0.000000 0.000000
Income            0.000000 0.000000 0.000000 0.000000
Outage_sec_perweek    0.000000 0.000000 0.000000 0.000000
Email             1.000000 0.000000 0.000000 0.000000
Contacts          0.000000 1.000000 0.000000 0.000000
Yearly_equip_failure    0.000000 0.000000 1.000000 0.000000
MonthlyCharge     0.000000 0.000000 0.000000 0.000000
Bandwidth_GB_Year    0.014844 0.000000 0.000000 0.000000
Bandwidth_GB_Year    0.002995 0.000000 0.000000 0.000000
Item1             0.000000 0.000000 0.000000 0.000000
Item2             0.000000 0.000000 0.000000 0.000000
Item3             0.000000 0.000000 0.000000 0.000000
Item4             0.014844 -0.000994 -0.000995 -0.000127
Item5             0.000000 0.000000 0.000000 0.000000
Item6             0.000000 0.000000 0.000000 0.000000
Item7             0.002204 -0.000000 0.000000 0.000000
Item8             0.000000 0.000000 0.000000 0.000000

                  MonthlyCharge  Bandwidth_GB_Year   Item1   Item2  \\
Customer_ID       0.000000 0.001967 0.000000 0.000000
Zip               -0.000000 0.724780 0.000000 0.000000
Lat               0.000000 -0.010845 0.000000 0.000000
Lng               0.000000 0.011765 0.000000 0.000000
Population        0.000000 0.000000 0.000000 0.000000
Children          0.000000 0.000000 0.000000 0.000000
Age               0.000000 0.000000 0.000000 0.000000
Income            0.000000 0.000000 0.000000 0.000000
Outage_sec_perweek    0.000000 0.000000 0.000000 0.000000
Email             0.000000 0.000000 0.000000 0.000000
Contacts          0.000000 0.000000 0.000000 0.000000
Yearly_equip_failure    0.000000 0.000000 0.000000 0.000000
MonthlyCharge     1.000000 0.000000 0.000000 0.000000
Bandwidth_GB_Year    0.000000 1.000000 0.000000 0.000000
Item1             0.000000 0.000000 1.000000 0.000000
Item2             0.000000 0.000000 0.000000 1.000000
Item3             0.000000 0.000000 0.000000 0.000000
Item4             0.000000 0.000000 0.000000 0.000000
Item5             0.000000 0.000000 0.000000 0.000000
Item6             0.000000 0.000000 0.000000 0.000000
Item7             0.000000 0.000000 0.000000 0.000000
Item8             0.000000 0.000000 0.000000 0.000000

                  Item3   Item4   Item5   Item6   Item7  \\
Customer_ID       0.002465 -0.006305 0.014429 -0.007937 0.002995
Zip               -0.968476 0.011111 -0.002609 0.000000 -5.537567e-03
Lat               0.003421 -0.010745 0.000000 0.000000 0.000897
Lng               0.003866 -0.007939 0.001364 -0.000778 0.003401
Population        0.000000 0.000000 0.000000 0.000000 0.000000
Children          0.000000 0.000000 0.000000 0.000000 0.000000
Age               0.000000 0.000000 0.000000 0.000000 0.000000
Income            0.000000 0.000000 0.000000 0.000000 0.000000
Outage_sec_perweek    0.000000 0.000000 0.000000 0.000000 0.000000
Email             0.000000 0.000000 0.000000 0.000000 0.000000
Contacts          0.000000 0.000000 0.000000 0.000000 0.000000
Yearly_equip_failure    0.000000 0.000000 0.000000 0.000000 0.000000
MonthlyCharge     0.000000 0.000000 0.000000 0.000000 0.000000
Bandwidth_GB_Year    0.000000 0.000000 0.000000 0.000000 0.000000
Item1             0.000000 0.000000 0.000000 0.000000 0.000000
Item2             0.000000 0.000000 0.000000 0.000000 0.000000
Item3             1.000000 -0.777706 -0.001362 0.000000 0.000000
Item4             -0.777706 1.000000 0.000000 0.000000 0.000000
Item5             -0.001362 0.000000 1.000000 -0.000000 0.000000
Item6             0.000000 0.000000 -0.000000 1.000000 0.000000
Item7             0.000000 0.000000 0.000000 0.000000 1.000000
Item8             0.000000 0.000000 0.000000 0.000000 0.000000

                  Item8
Customer_ID       0.014840
Zip               -0.037761
Lat               0.000000
Lng               0.000000
Population        0.000000
Children          0.000000
Age               0.000000
Income            0.000000
Outage_sec_perweek    0.000000
Email             0.000000
Contacts          0.000000
Yearly_equip_failure    0.000000
MonthlyCharge     0.000000
Bandwidth_GB_Year    0.000000
Item1             0.000000
Item2             0.000000
Item3             0.000000
Item4             0.000000
Item5             0.000000
Item6             0.000000
Item7             0.000000
Item8             1.000000
```

```
In [103]: '''Customer and Bandwidth_GB_Year have been identified as having the strongest relationship with Tenure
          Bandwidth_GB_Year is the only candidate to include in a second model based on correlation of 0.825561
          Further exploration of the original data set is required for this process and a correlation matrix to assess
          additional predictors will be developed to this additional analysis.'''
```

```
Out[103]: 'Customer and Bandwidth_GB_Year have been identified as having the strongest relationship with TenureBandwidth_GB_Year is the only candidate to include in a second m
          odel based on correlation of 0.825561Further exploration of the original data set is required for this process and a correlation matrix to assess
          additional predictors will be developed.\n\n\n
```