Code

December 9, 2022

[ ]:
```
#importing req. Lib.
import pandas as pd
import numpy as np
import seaborn as sn
import matplotlib.pyplot as plt
import re
import nltk
from nltk.corpus import stopwords
from sklearn.model_selection import train_test_split
from mlxtend.plotting import plot_confusion_matrix
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import␣
 ↪accuracy_score,confusion_matrix,classification_report
from wordcloud import WordCloud
```

[ ]:
```
data = pd.read_csv('datasetPython.csv')
```

[ ]:
```
#data.head()
data.columns
data.info()

# data.nunique()

# data.isnull().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40102 entries, 0 to 40101
Data columns (total 29 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   ID                       40102 non-null  int64
 1   Language                 40102 non-null  object
 2   Subscription Type        40102 non-null  object
 3   Subscription Event Type  40102 non-null  object
 4   Purchase Store           40102 non-null  object
 5   Purchase Amount          25294 non-null  float64
 6   Currency                 26924 non-null  object
```

```
 7   Subscription Start Date  40102 non-null  object
 8   Subscription Expiration  40102 non-null  object
 9   Demo User                40102 non-null  object
 10  Free Trial User          40102 non-null  object
 11  Free Trial Start Date    5833 non-null   object
 12  Free Trial Expiration    5833 non-null   object
 13  Auto Renew               40101 non-null  object
 14  Country                  40102 non-null  object
 15  User Type                40102 non-null  object
 16  Lead Platform            40102 non-null  object
 17  Email Subscriber         40102 non-null  object
 18  Push Notifications       40102 non-null  object
 19  Send Count               28448 non-null  float64
 20  Open Count               28448 non-null  float64
 21  Click Count              28448 non-null  float64
 22  Unique Open Count        28448 non-null  float64
 23  Unique Click Count       28448 non-null  float64
 24  Sub Start Date Out       40102 non-null  object
 25  Sub Expiration Date Out  40102 non-null  object
 26  Subscription Length      40102 non-null  int64
 27  Send to Open Metric      28448 non-null  float64
 28  Open to Click Metric     15086 non-null  float64
dtypes: float64(8), int64(2), object(19)
memory usage: 8.9+ MB
```

[ ]:
```
# data.isnull().sum().plot(kind='bar')
```

[ ]:
```
# Converting the dates to datetime format for further analysis
data[['Subscription Start Date', 'Subscription Expiration', 'Free Trial Start␣
 ↪Date', 'Free Trial Expiration', 'Sub Start Date Out', 'Sub Expiration Date␣
 ↪Out']] = data[['Subscription Start Date', 'Subscription Expiration', 'Free␣
 ↪Trial Start Date', 'Free Trial Expiration', 'Sub Start Date Out', 'Sub␣
 ↪Expiration Date Out']].astype('datetime64[ns]')
```

[ ]:
```
# Double checking that conversion of type worked. It did
data.info()
```

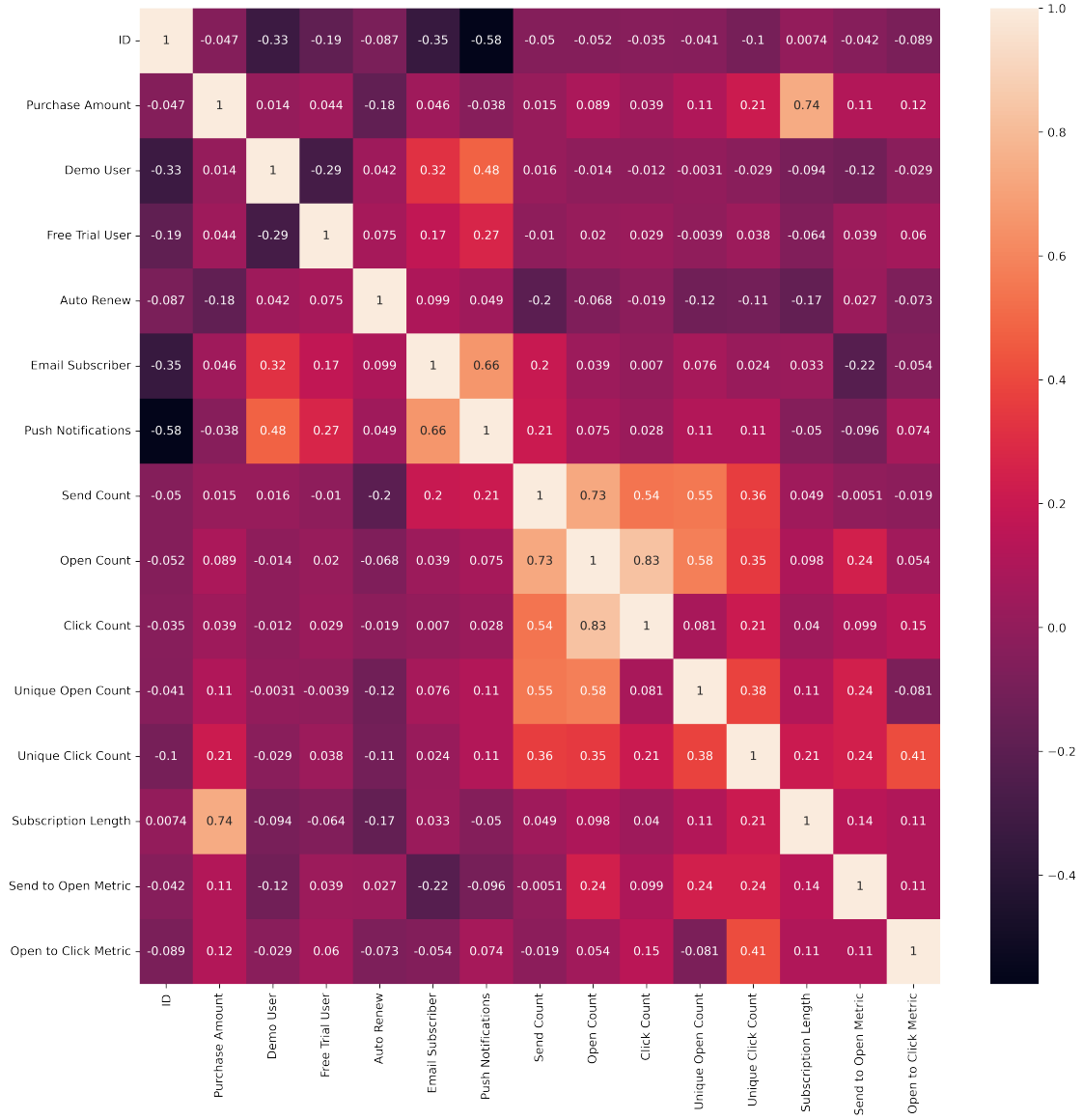
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40102 entries, 0 to 40101
Data columns (total 29 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   ID                       40102 non-null  int64
 1   Language                 40102 non-null  object
 2   Subscription Type        40102 non-null  object
 3   Subscription Event Type  40102 non-null  object
 4   Purchase Store           40102 non-null  object
 5   Purchase Amount          25294 non-null  float64
```

```
 6   Currency                 26924 non-null  object
 7   Subscription Start Date  40102 non-null  datetime64[ns]
 8   Subscription Expiration  40102 non-null  datetime64[ns]
 9   Demo User                40102 non-null  object
 10  Free Trial User          40102 non-null  object
 11  Free Trial Start Date    5833 non-null   datetime64[ns]
 12  Free Trial Expiration    5833 non-null   datetime64[ns]
 13  Auto Renew               40101 non-null  object
 14  Country                  40102 non-null  object
 15  User Type                40102 non-null  object
 16  Lead Platform            40102 non-null  object
 17  Email Subscriber         40102 non-null  object
 18  Push Notifications       40102 non-null  object
 19  Send Count               28448 non-null  float64
 20  Open Count               28448 non-null  float64
 21  Click Count              28448 non-null  float64
 22  Unique Open Count        28448 non-null  float64
 23  Unique Click Count       28448 non-null  float64
 24  Sub Start Date Out       40102 non-null  datetime64[ns]
 25  Sub Expiration Date Out  40102 non-null  datetime64[ns]
 26  Subscription Length      40102 non-null  int64
 27  Send to Open Metric      28448 non-null  float64
 28  Open to Click Metric     15086 non-null  float64
dtypes: datetime64[ns](6), float64(8), int64(2), object(13)
memory usage: 8.9+ MB
```

[ ]:
```
# Convert data types to 1 and 0 for easier analysis
data['Demo User'] = data['Demo User'].map({'Yes': 1, 'No': 0})
data['Free Trial User'] = data['Free Trial User'].map({'Yes': 1, 'No': 0})
data['Auto Renew'] = data['Auto Renew'].map({'On': 1, 'Off': 0})
data['Email Subscriber'] = data['Email Subscriber'].map({'Yes': 1, 'No': 0})
data['Push Notifications'] = data['Push Notifications'].map({'Yes': 1, 'No': 0})
```

[ ]:
```
data.info()
```

[ ]:
```
fig = plt.figure(figsize=(15, 15), dpi=480)
corr_matrix = data.corr()
# print(corr_matrix)
sn.heatmap(corr_matrix, annot=True)
plt.show()
```



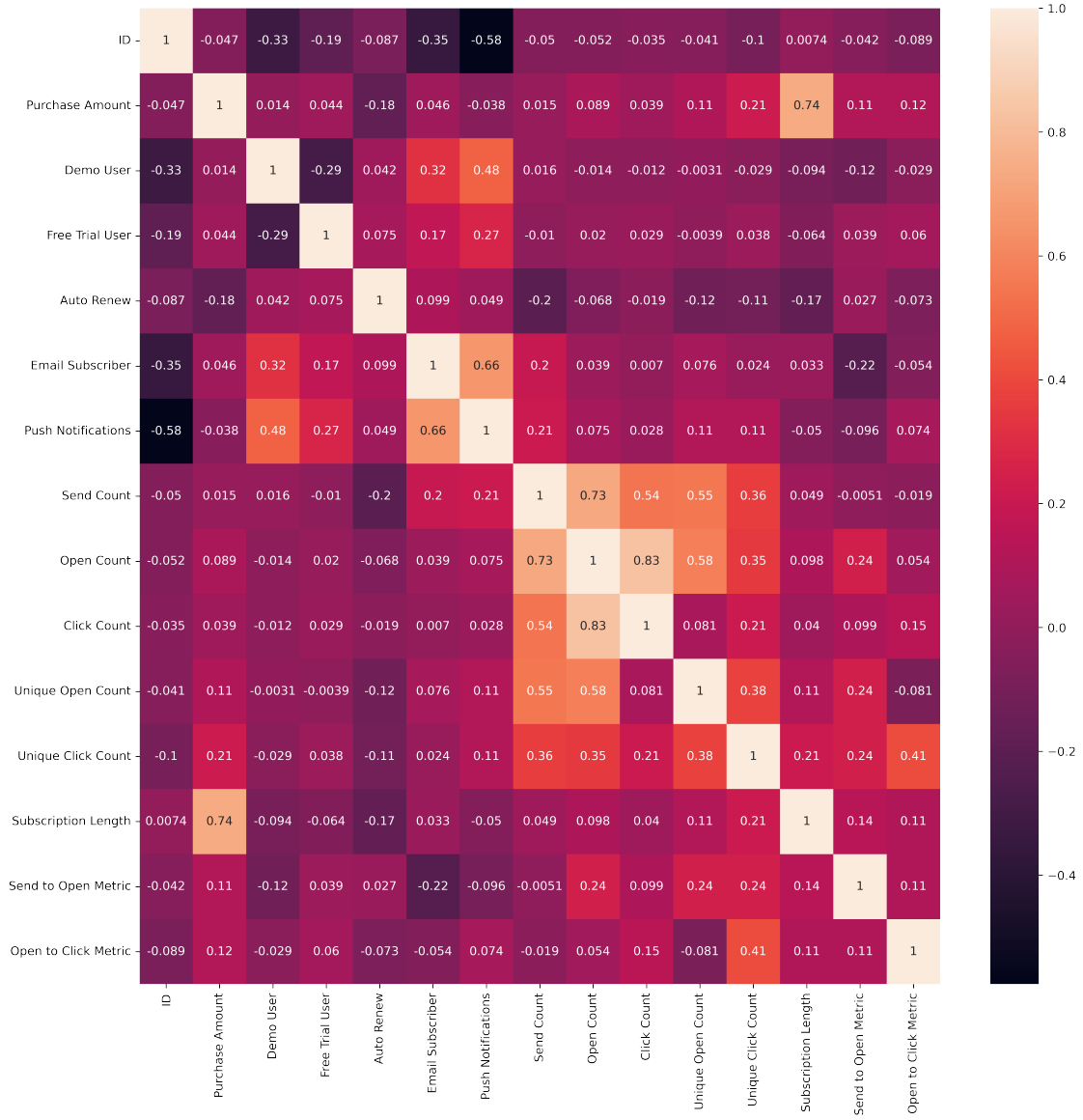
[ ]:
```
# # Converting more columns from dtype object to dtype category
data[['Language', 'Subscription Type', 'Subscription Event Type', 'Purchase␣
 ↪Store', 'Currency', 'Country', 'User Type', 'Lead Platform',]] =␣
 ↪data[['Language', 'Subscription Type', 'Subscription Event Type', 'Purchase␣
 ↪Store', 'Currency', 'Country', 'User Type', 'Lead Platform']].
 ↪astype('category')
```

[ ]:
```
# Double checking that conversion of type worked. It did!
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40102 entries, 0 to 40101
```

```
Data columns (total 29 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   ID                       40102 non-null  int64
 1   Language                 40102 non-null  category
 2   Subscription Type        40102 non-null  category
 3   Subscription Event Type  40102 non-null  category
 4   Purchase Store           40102 non-null  category
 5   Purchase Amount          25294 non-null  float64
 6   Currency                 26924 non-null  category
 7   Subscription Start Date  40102 non-null  datetime64[ns]
 8   Subscription Expiration  40102 non-null  datetime64[ns]
 9   Demo User                40102 non-null  int64
 10  Free Trial User          40102 non-null  int64
 11  Free Trial Start Date    5833 non-null   datetime64[ns]
 12  Free Trial Expiration    5833 non-null   datetime64[ns]
 13  Auto Renew               40101 non-null  float64
 14  Country                  40102 non-null  category
 15  User Type                40102 non-null  category
 16  Lead Platform            40102 non-null  category
 17  Email Subscriber         40102 non-null  int64
 18  Push Notifications       40102 non-null  int64
 19  Send Count               28448 non-null  float64
 20  Open Count               28448 non-null  float64
 21  Click Count              28448 non-null  float64
 22  Unique Open Count        28448 non-null  float64
 23  Unique Click Count       28448 non-null  float64
 24  Sub Start Date Out       40102 non-null  datetime64[ns]
 25  Sub Expiration Date Out  40102 non-null  datetime64[ns]
 26  Subscription Length      40102 non-null  int64
 27  Send to Open Metric      28448 non-null  float64
 28  Open to Click Metric     15086 non-null  float64
dtypes: category(8), datetime64[ns](6), float64(9), int64(6)
memory usage: 6.7 MB
```

[ ]:
```python
# New correlation matrix after converting columns to dtype category
fig = plt.figure(figsize=(15, 15), dpi=480)
corr_matrix = data.corr()
# print(corr_matrix)
sn.heatmap(corr_matrix, annot=True)
plt.show()
```



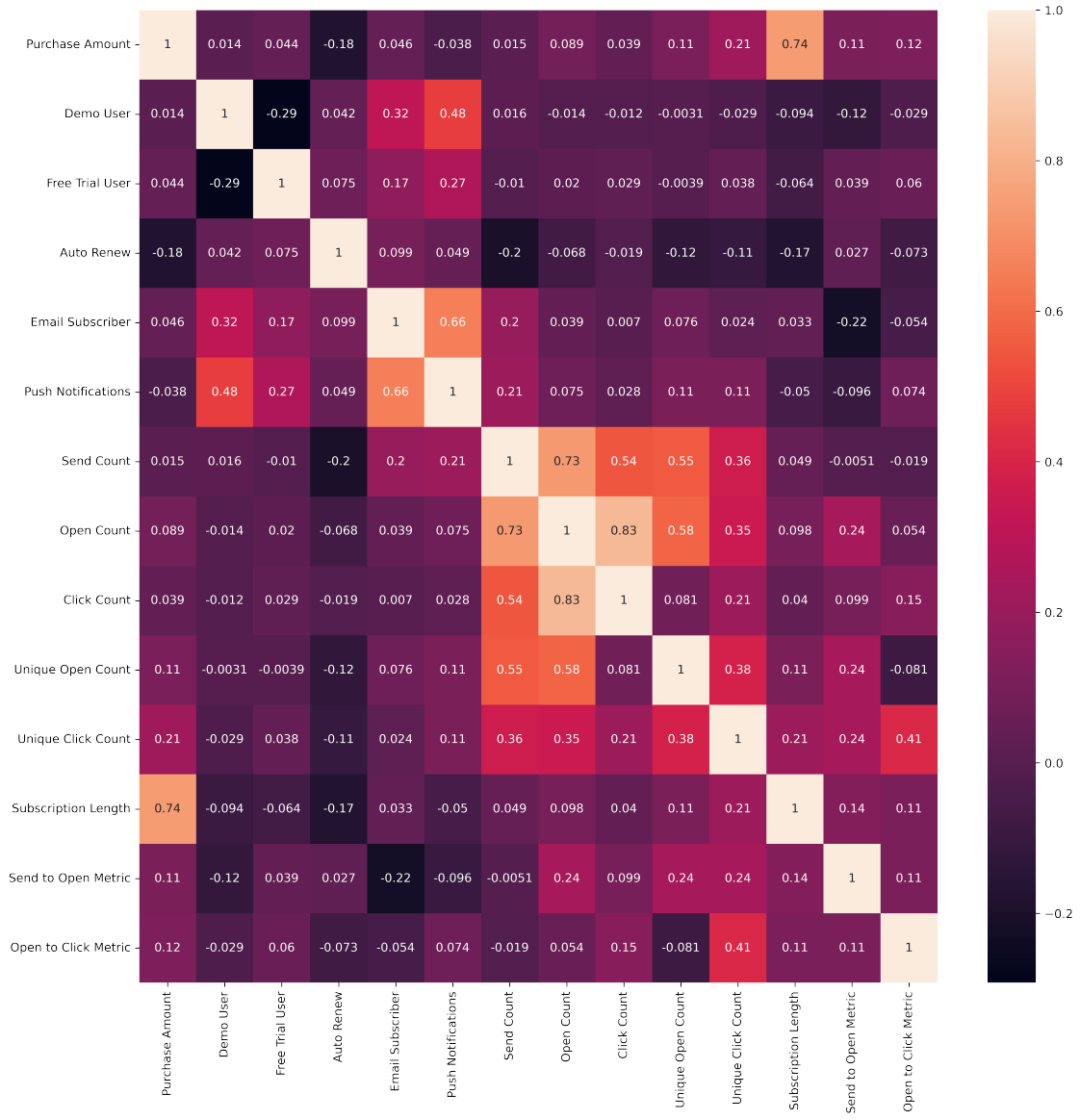
[ ]:
```
# Time to drop our unnecessary columns
# We'll start with currency as the purchase amounts have all been converted to␣
↪USD
data.drop('Currency', axis=1, inplace=True)
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40102 entries, 0 to 40101
Data columns (total 28 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   ID                      40102 non-null  int64
```

```
 1   Language                 40102 non-null  category
 2   Subscription Type        40102 non-null  category
 3   Subscription Event Type  40102 non-null  category
 4   Purchase Store           40102 non-null  category
 5   Purchase Amount          25294 non-null  float64
 6   Subscription Start Date  40102 non-null  datetime64[ns]
 7   Subscription Expiration  40102 non-null  datetime64[ns]
 8   Demo User                40102 non-null  int64
 9   Free Trial User          40102 non-null  int64
 10  Free Trial Start Date    5833 non-null   datetime64[ns]
 11  Free Trial Expiration    5833 non-null   datetime64[ns]
 12  Auto Renew               40101 non-null  float64
 13  Country                  40102 non-null  category
 14  User Type                40102 non-null  category
 15  Lead Platform            40102 non-null  category
 16  Email Subscriber         40102 non-null  int64
 17  Push Notifications       40102 non-null  int64
 18  Send Count               28448 non-null  float64
 19  Open Count               28448 non-null  float64
 20  Click Count              28448 non-null  float64
 21  Unique Open Count        28448 non-null  float64
 22  Unique Click Count       28448 non-null  float64
 23  Sub Start Date Out       40102 non-null  datetime64[ns]
 24  Sub Expiration Date Out  40102 non-null  datetime64[ns]
 25  Subscription Length      40102 non-null  int64
 26  Send to Open Metric      28448 non-null  float64
 27  Open to Click Metric     15086 non-null  float64
dtypes: category(7), datetime64[ns](6), float64(9), int64(6)
memory usage: 6.7 MB
```

[ ]:
```
# updating the data frame to remove ID as we don't want it to impact our model␣
↪and analysis
newData = data.drop('ID', axis=1)
# newData.info()

# New correlation matrix after dropping ID column
fig = plt.figure(figsize=(15, 15), dpi=480)
corr_matrix = newData.corr()
# print(corr_matrix)
sn.heatmap(corr_matrix, annot=True)
plt.show()
```



[ ]:
```
dfDummies = pd.get_dummies(newData, columns=['Subscription Type', 'Subscription␣
↪Event Type', 'Purchase Store', 'Country', 'User Type', 'Lead Platform'])
dfDummies2 = pd.get_dummies(newData, columns=['Language', 'Subscription Type',␣
↪'Subscription Event Type', 'Purchase Store', 'Country', 'User Type', 'Lead␣
↪Platform'])
```

[ ]:
```
dfDummies.info()
dfDummies2.info()
```

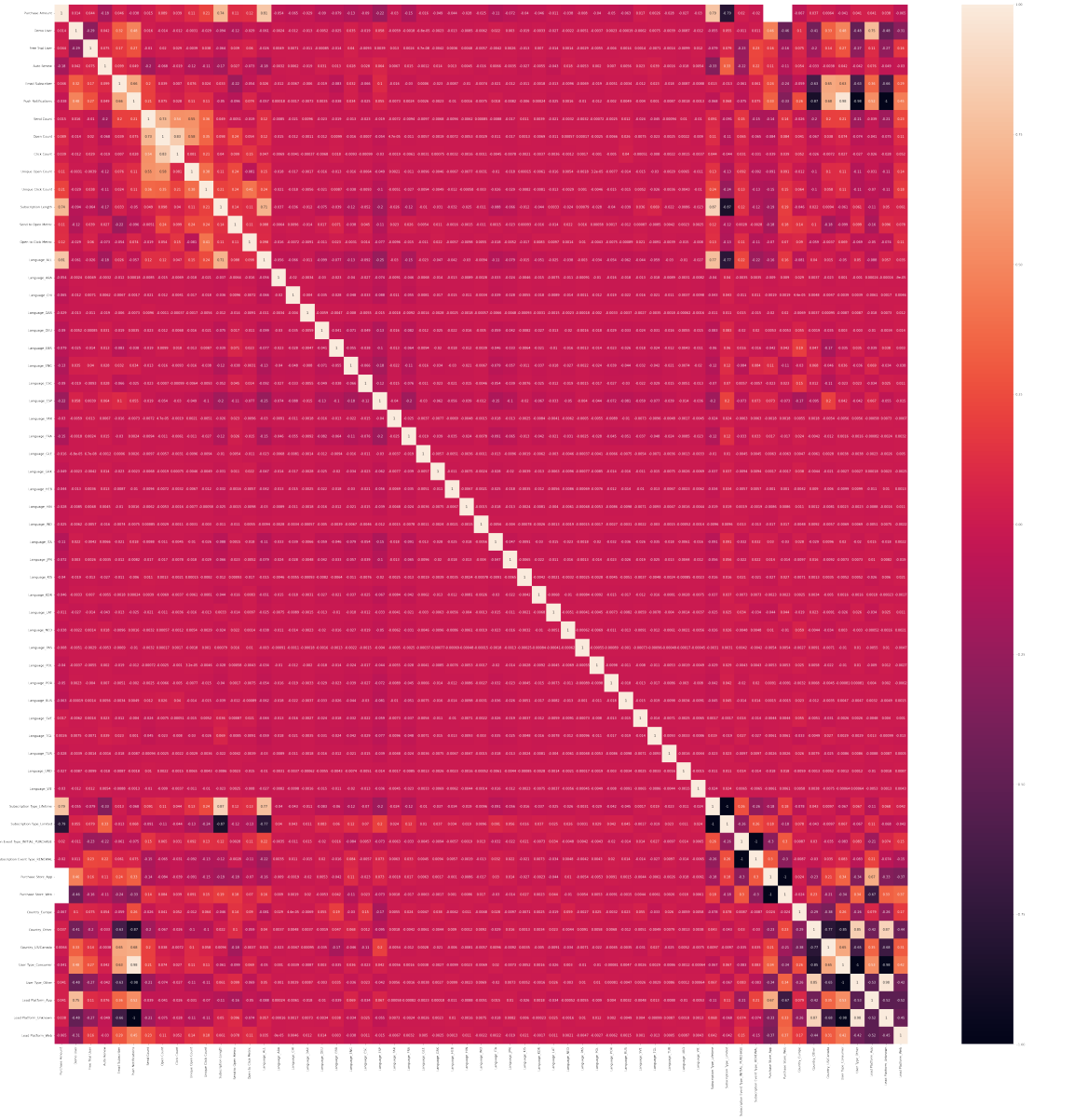
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40102 entries, 0 to 40101
Data columns (total 35 columns):
```

```
 #   Column                                    Non-Null Count  Dtype
---  ------                                    --------------  -----
 0   Language                                  40102 non-null  category
 1   Purchase Amount                           25294 non-null  float64
 2   Subscription Start Date                   40102 non-null  datetime64[ns]
 3   Subscription Expiration                   40102 non-null  datetime64[ns]
 4   Demo User                                 40102 non-null  int64
 5   Free Trial User                           40102 non-null  int64
 6   Free Trial Start Date                     5833 non-null   datetime64[ns]
 7   Free Trial Expiration                     5833 non-null   datetime64[ns]
 8   Auto Renew                                40101 non-null  float64
 9   Email Subscriber                          40102 non-null  int64
 10  Push Notifications                        40102 non-null  int64
 11  Send Count                                28448 non-null  float64
 12  Open Count                                28448 non-null  float64
 13  Click Count                               28448 non-null  float64
 14  Unique Open Count                         28448 non-null  float64
 15  Unique Click Count                        28448 non-null  float64
 16  Sub Start Date Out                        40102 non-null  datetime64[ns]
 17  Sub Expiration Date Out                   40102 non-null  datetime64[ns]
 18  Subscription Length                       40102 non-null  int64
 19  Send to Open Metric                       28448 non-null  float64
 20  Open to Click Metric                      15086 non-null  float64
 21  Subscription Type_Lifetime                40102 non-null  uint8
 22  Subscription Type_Limited                 40102 non-null  uint8
 23  Subscription Event Type_INITIAL_PURCHASE  40102 non-null  uint8
 24  Subscription Event Type_RENEWAL           40102 non-null  uint8
 25  Purchase Store_App                        40102 non-null  uint8
 26  Purchase Store_Web                        40102 non-null  uint8
 27  Country_Europe                            40102 non-null  uint8
 28  Country_Other                             40102 non-null  uint8
 29  Country_US/Canada                         40102 non-null  uint8
 30  User Type_Consumer                        40102 non-null  uint8
 31  User Type_Other                           40102 non-null  uint8
 32  Lead Platform_App                         40102 non-null  uint8
 33  Lead Platform_Unknown                     40102 non-null  uint8
 34  Lead Platform_Web                         40102 non-null  uint8
dtypes: category(1), datetime64[ns](6), float64(9), int64(5), uint8(14)
memory usage: 6.7 MB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40102 entries, 0 to 40101
Data columns (total 65 columns):
 #   Column                                    Non-Null Count  Dtype
---  ------                                    --------------  -----
 0   Purchase Amount                           25294 non-null  float64
 1   Subscription Start Date                   40102 non-null  datetime64[ns]
 2   Subscription Expiration                   40102 non-null  datetime64[ns]
 3   Demo User                                 40102 non-null  int64
```

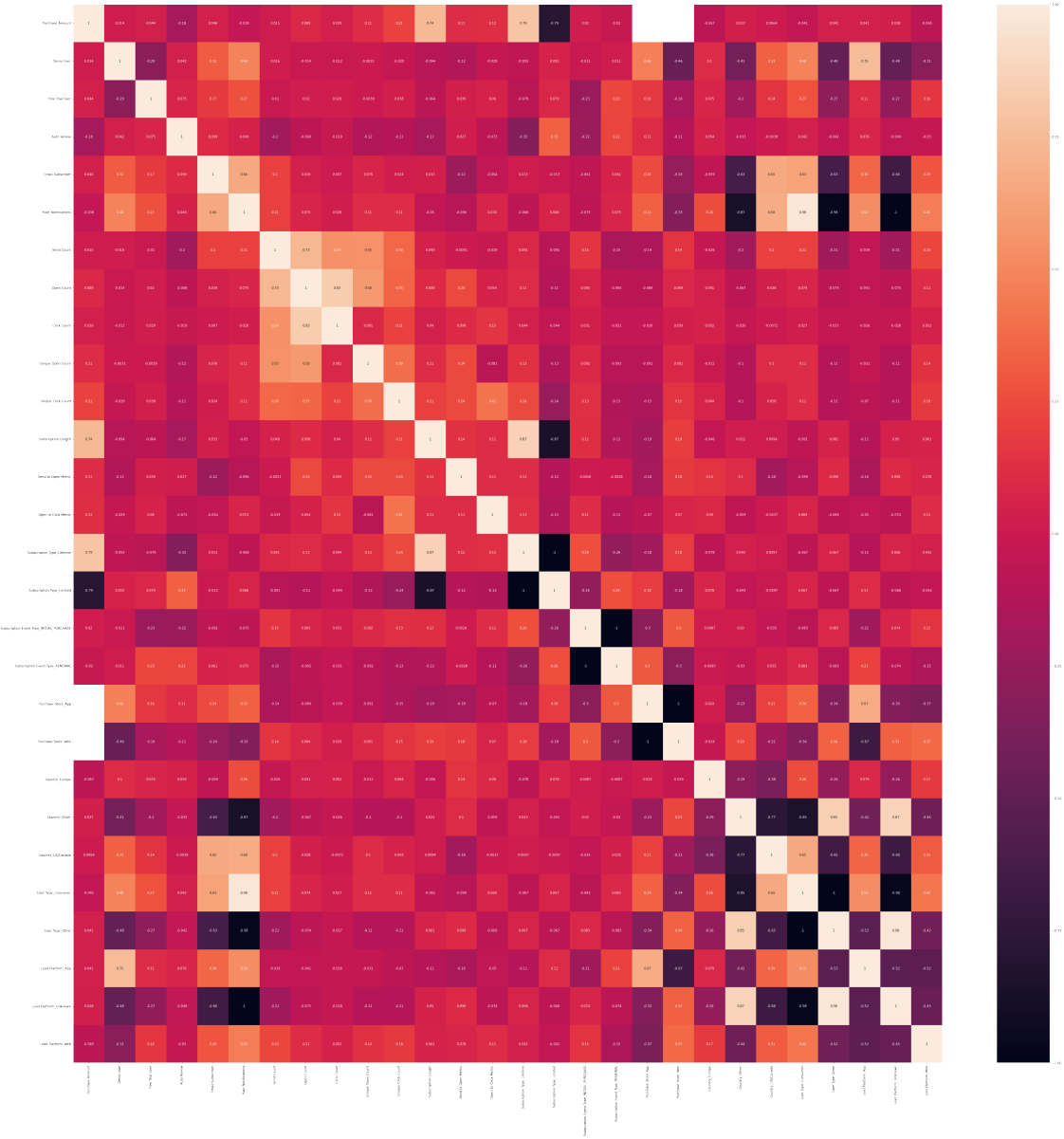
```
4   Free Trial User                  40102 non-null  int64
5   Free Trial Start Date            5833 non-null   datetime64[ns]
6   Free Trial Expiration            5833 non-null   datetime64[ns]
7   Auto Renew                       40101 non-null  float64
8   Email Subscriber                 40102 non-null  int64
9   Push Notifications               40102 non-null  int64
10  Send Count                       28448 non-null  float64
11  Open Count                       28448 non-null  float64
12  Click Count                      28448 non-null  float64
13  Unique Open Count                28448 non-null  float64
14  Unique Click Count               28448 non-null  float64
15  Sub Start Date Out               40102 non-null  datetime64[ns]
16  Sub Expiration Date Out          40102 non-null  datetime64[ns]
17  Subscription Length              40102 non-null  int64
18  Send to Open Metric              28448 non-null  float64
19  Open to Click Metric             15086 non-null  float64
20  Language_ALL                     40102 non-null  uint8
21  Language_ARA                     40102 non-null  uint8
22  Language_CHI                     40102 non-null  uint8
23  Language_DAR                     40102 non-null  uint8
24  Language_DEU                     40102 non-null  uint8
25  Language_EBR                     40102 non-null  uint8
26  Language_ENG                     40102 non-null  uint8
27  Language_ESC                     40102 non-null  uint8
28  Language_ESP                     40102 non-null  uint8
29  Language_FAR                     40102 non-null  uint8
30  Language_FRA                     40102 non-null  uint8
31  Language_GLE                     40102 non-null  uint8
32  Language_GRK                     40102 non-null  uint8
33  Language_HEB                     40102 non-null  uint8
34  Language_HIN                     40102 non-null  uint8
35  Language_IND                     40102 non-null  uint8
36  Language_ITA                     40102 non-null  uint8
37  Language_JPN                     40102 non-null  uint8
38  Language_KIS                     40102 non-null  uint8
39  Language_KOR                     40102 non-null  uint8
40  Language_LAT                     40102 non-null  uint8
41  Language_NED                     40102 non-null  uint8
42  Language_PAS                     40102 non-null  uint8
43  Language_POL                     40102 non-null  uint8
44  Language_POR                     40102 non-null  uint8
45  Language_RUS                     40102 non-null  uint8
46  Language_SVE                     40102 non-null  uint8
47  Language_TGL                     40102 non-null  uint8
48  Language_TUR                     40102 non-null  uint8
49  Language_URD                     40102 non-null  uint8
50  Language_VIE                     40102 non-null  uint8
51  Subscription Type_Lifetime       40102 non-null  uint8
```

```
 52  Subscription Type_Limited                 40102 non-null  uint8
 53  Subscription Event Type_INITIAL_PURCHASE  40102 non-null  uint8
 54  Subscription Event Type_RENEWAL           40102 non-null  uint8
 55  Purchase Store_App                        40102 non-null  uint8
 56  Purchase Store_Web                        40102 non-null  uint8
 57  Country_Europe                            40102 non-null  uint8
 58  Country_Other                             40102 non-null  uint8
 59  Country_US/Canada                         40102 non-null  uint8
 60  User Type_Consumer                        40102 non-null  uint8
 61  User Type_Other                           40102 non-null  uint8
 62  Lead Platform_App                         40102 non-null  uint8
 63  Lead Platform_Unknown                     40102 non-null  uint8
 64  Lead Platform_Web                         40102 non-null  uint8
dtypes: datetime64[ns](6), float64(9), int64(5), uint8(45)
memory usage: 7.8 MB
```

[ ]:
```
# New correlation matrix after making dummies variables for correlation matrix
fig = plt.figure(figsize=(60, 60)) # , dpi=480)
corr_matrix = dfDummies2.corr()
#print(corr_matrix)
sn.heatmap(corr_matrix, annot=True)
plt.show()
```



[ ]:
```
# New correlation matrix after making dummies variables for correlation matrix
fig = plt.figure(figsize=(60, 60)) # , dpi=480)
corr_matrix = dfDummies.corr()
#print(corr_matrix)
sn.heatmap(corr_matrix, annot=True)
plt.show()
```



[ ]:
```
dfDummies3 = dfDummies.drop(['Subscription Length'], axis=1)

# New correlation matrix after making dummies variables for correlation matrix
fig = plt.figure(figsize=(60, 60)) # , dpi=480)
corr_matrix = dfDummies3.corr()
#print(corr_matrix)
sn.heatmap(corr_matrix, annot=True)
plt.show()
```