By: Brian Kim

In [1]:
```python
import numpy as np
import pandas as pd
from plotnine import *

import warnings
warnings.filterwarnings('ignore')
```

Web Dummy Variables

In [2]:
```python
data = pd.read_csv("Subscriber Information (Clean) Version 4.5.csv")
```

In [3]:
```python
del data['Subscription_Start_Date']
del data['Subscription_Expiration']
```

In [4]:
```python
data.info()
```

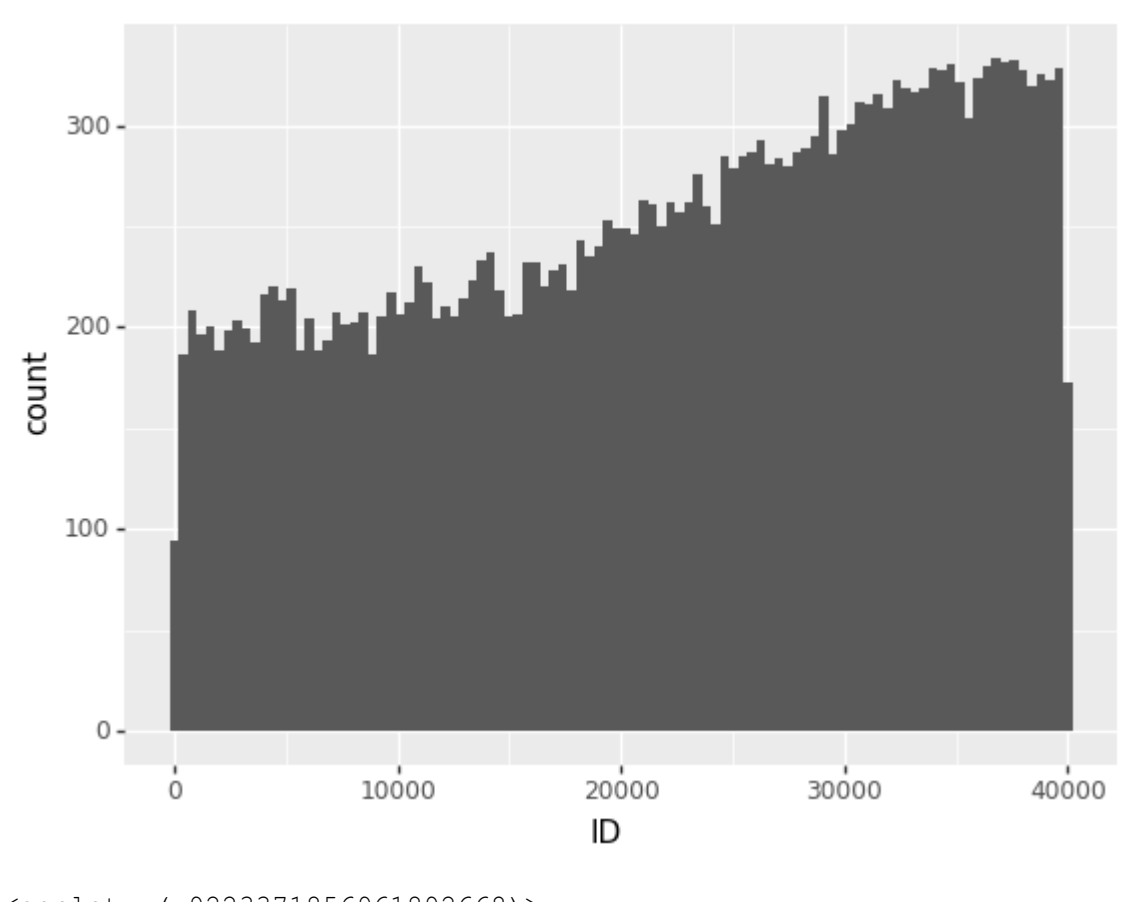
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25291 entries, 0 to 25290
Data columns (total 27 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   ID                       25291 non-null  int64  
 1   Language                 25291 non-null  object 
 2   Subscription_Type        25291 non-null  object 
 3   Subscription_Event_Type  25291 non-null  object 
 4   Purchase_Store           25291 non-null  object 
 5   Purchase_Amounts         25291 non-null  float64
 6   Duration                 25291 non-null  int64  
 7   Demo_User                25291 non-null  object 
 8   Free_Trial_User          25291 non-null  object 
 9   Auto_Renew               25291 non-null  object 
 10  Country                  25291 non-null  object 
 11  User_Type                25291 non-null  object 
 12  Lead_Platform            25291 non-null  object 
 13  Email_Subscriber         25291 non-null  object 
 14  Push_Notifications       25291 non-null  object 
 15  Send_Count               25291 non-null  int64  
 16  Open_Count               25291 non-null  int64  
 17  Click_Count              25291 non-null  int64  
 18  Unique_Open_Count        25291 non-null  int64  
 19  Unique_Click_Count       25291 non-null  int64  
 20  Start                    25291 non-null  int64  
 21  Other                    25291 non-null  int64  
 22  Completed                25291 non-null  int64  
 23  NULL                     25291 non-null  int64  
 24  Onboarding               25291 non-null  int64  
 25  App_Launch_Times         25291 non-null  int64  
 26  Total_Time_Launched      25291 non-null  int64  
dtypes: float64(1), int64(14), object(12)
memory usage: 5.2+ MB
```

Numerical

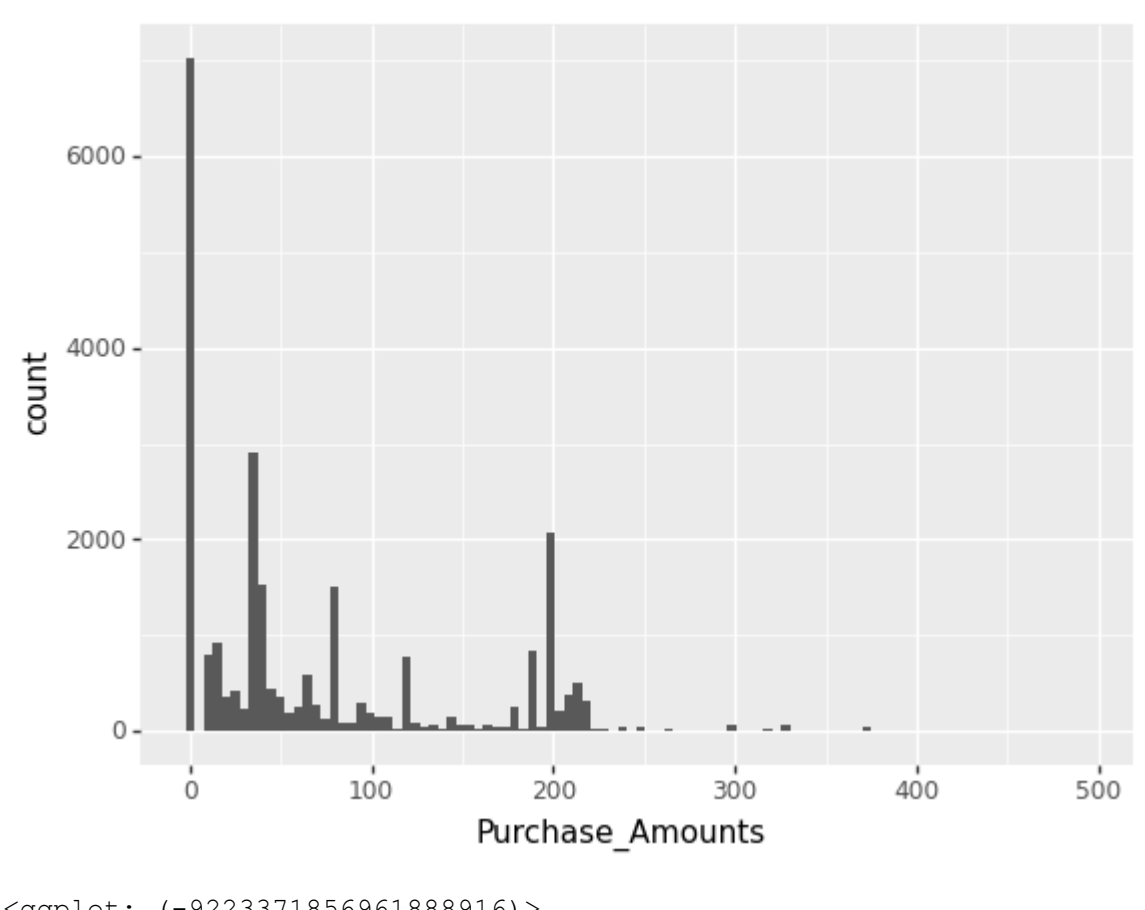
- ID
- Purchase_Amounts
- Duration
- Send_Count
- Open_Count
- Click_Count
- Unique_Open_Count
- Unique_Click_Count
- Start
- Other
- Completed
- NULL
- Onboarding
- App_Launch_Times
- Total_Time_Launched

In [5]:
```python
numerical_columns = ['ID', 'Purchase_Amounts', 'Duration', 'Send_Count', 'Open_Count', 'Click_Count',
'Unique_Open_Count', 'Unique_Click_Count', 'Start', 'Other', 'Completed', 'NULL', 'Onboarding', 'App_La
unch_Times', 'Total_Time_Launched']
numerical = data[numerical_columns]
```

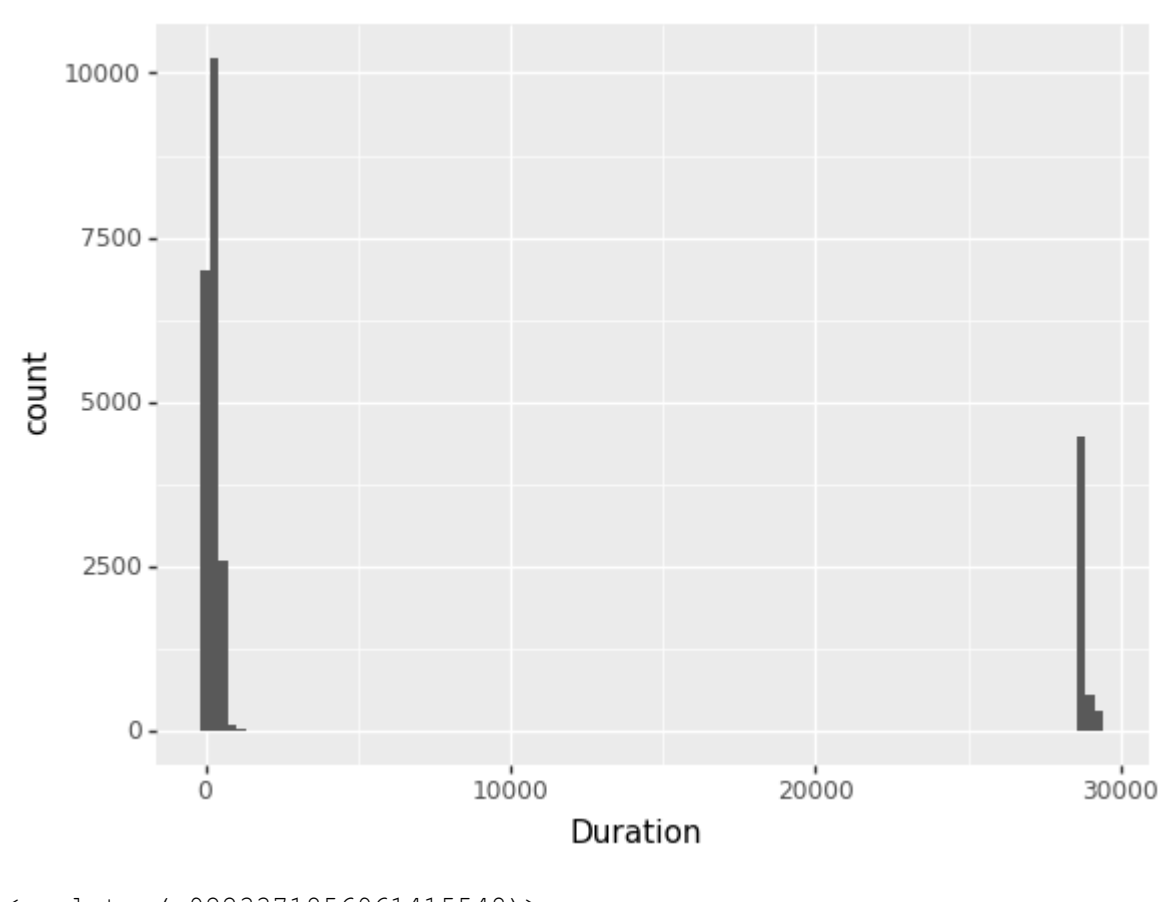
In [6]:
```python
for i in numerical_columns:
    print(ggplot(numerical, aes(x = i)) + geom_histogram(bins = 100))
```



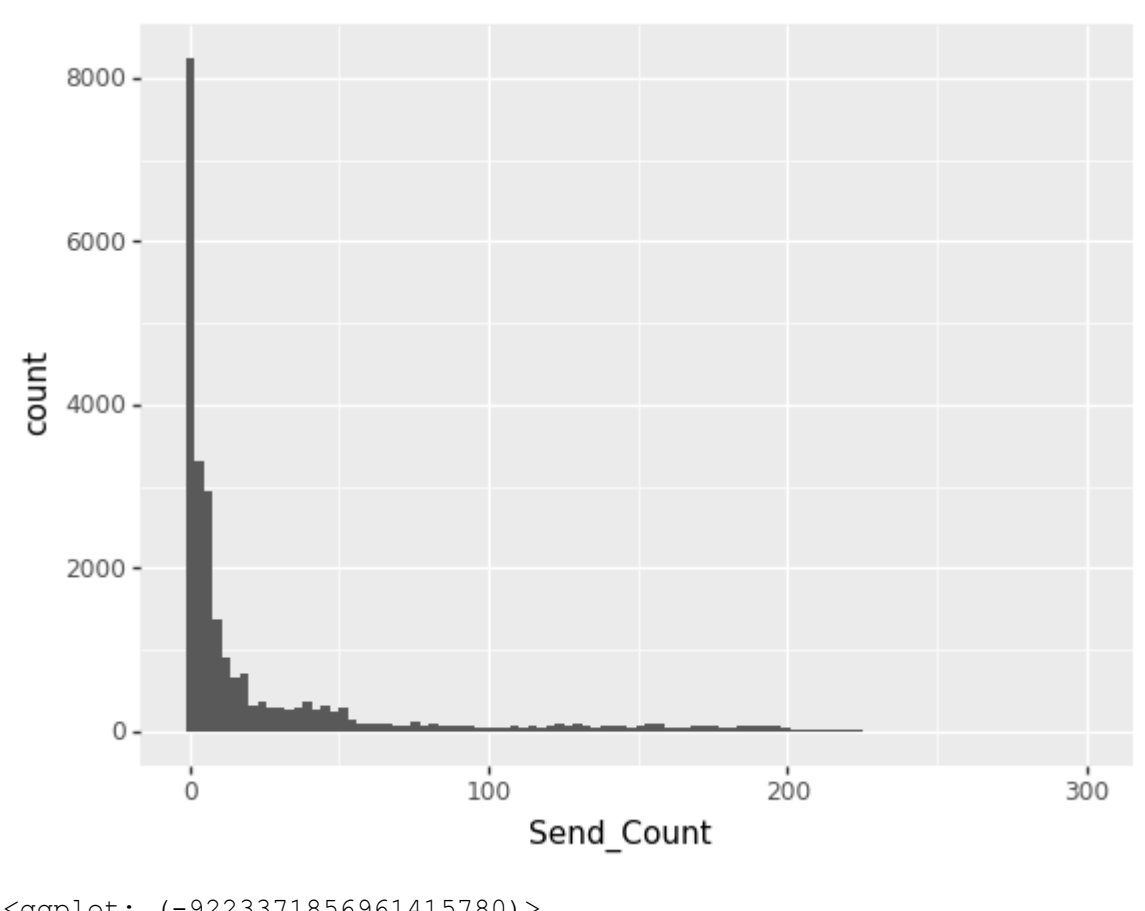
<ggplot: (-9223371856961892668)>



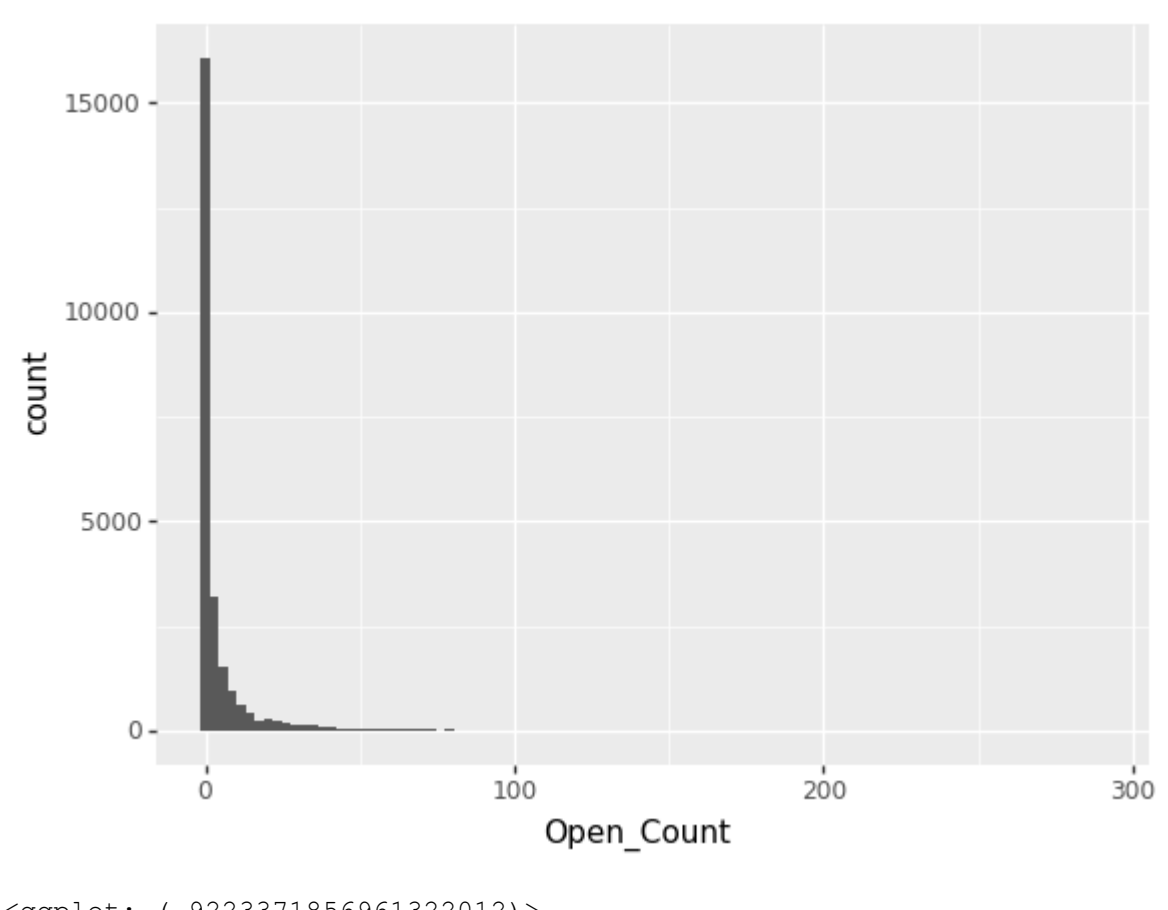
<ggplot: (-9223371856961888916)>



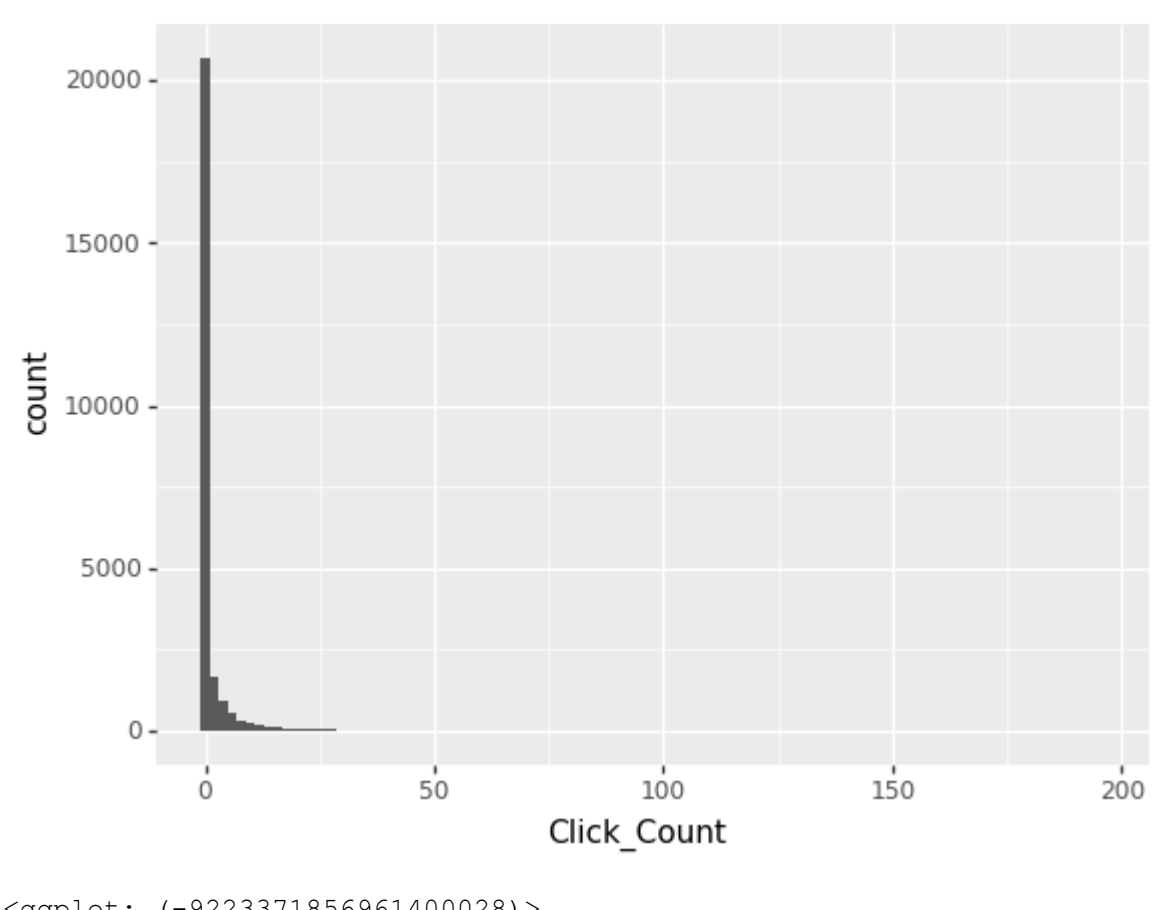
<ggplot: (-9223371856961415548)>



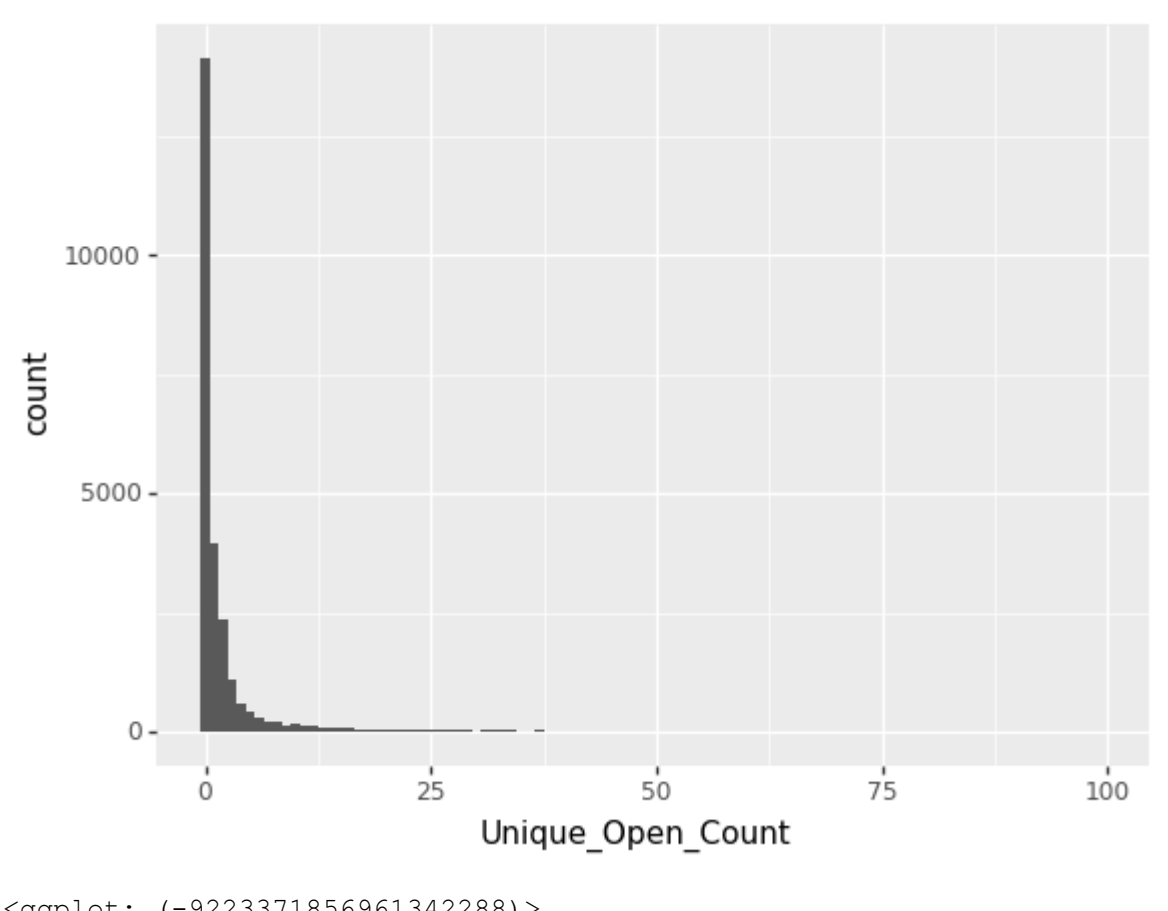
<ggplot: (-9223371856961415780)>



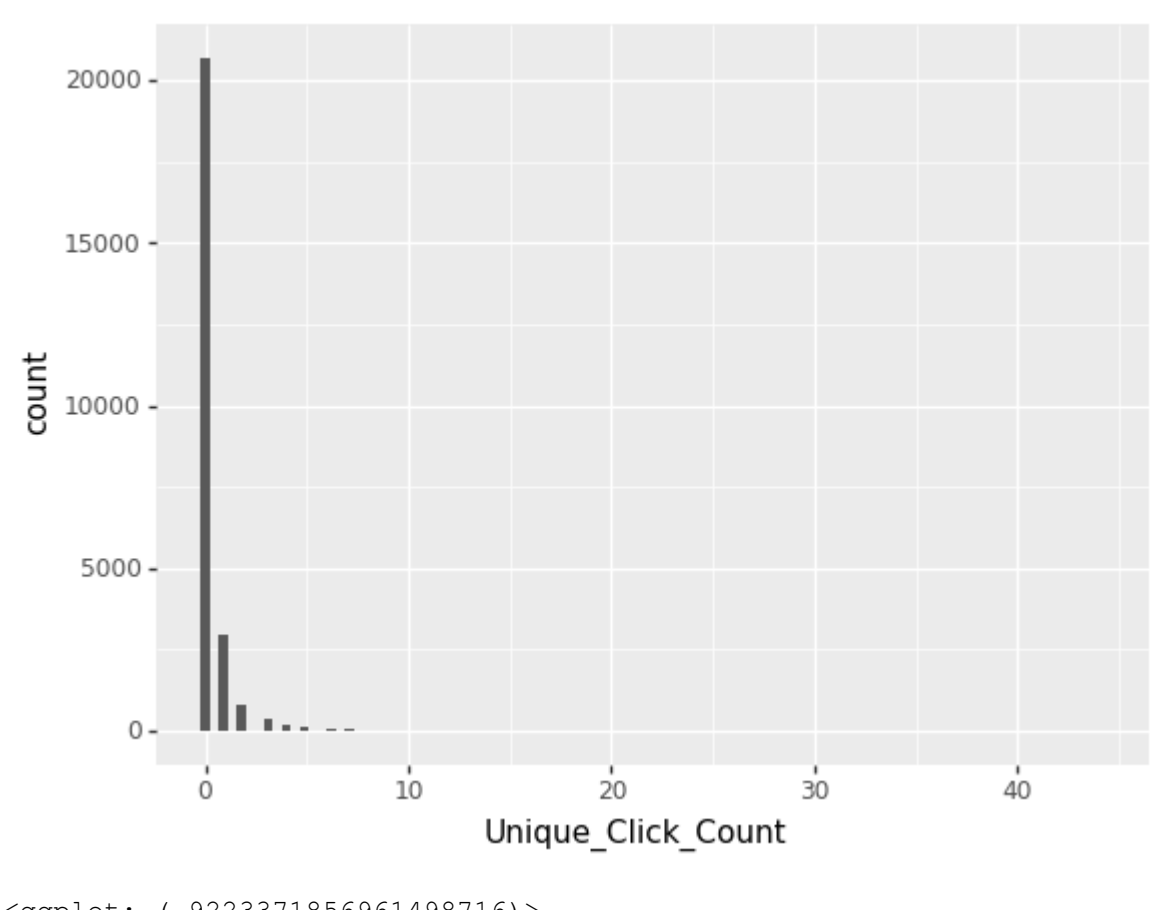
<ggplot: (-9223371856961322012)>



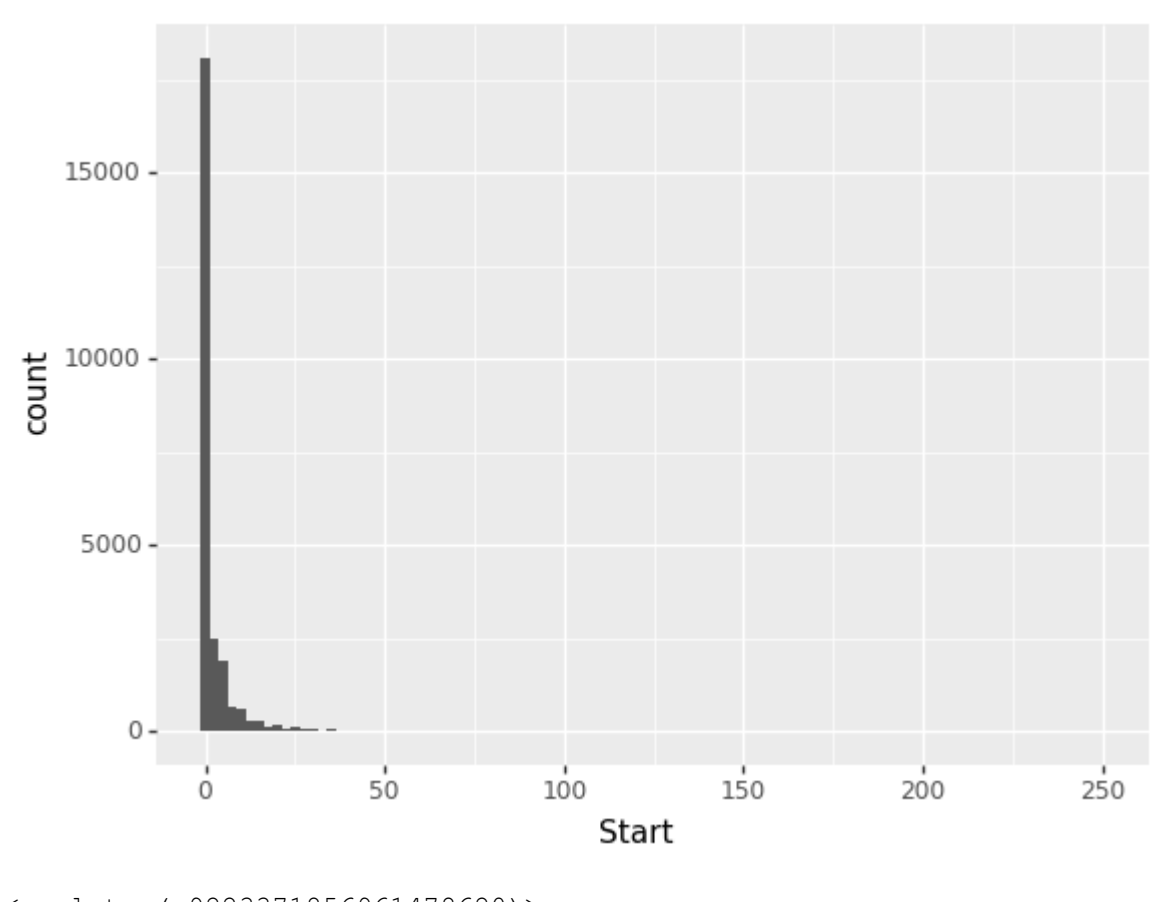
<ggplot: (-9223371856961400028)>



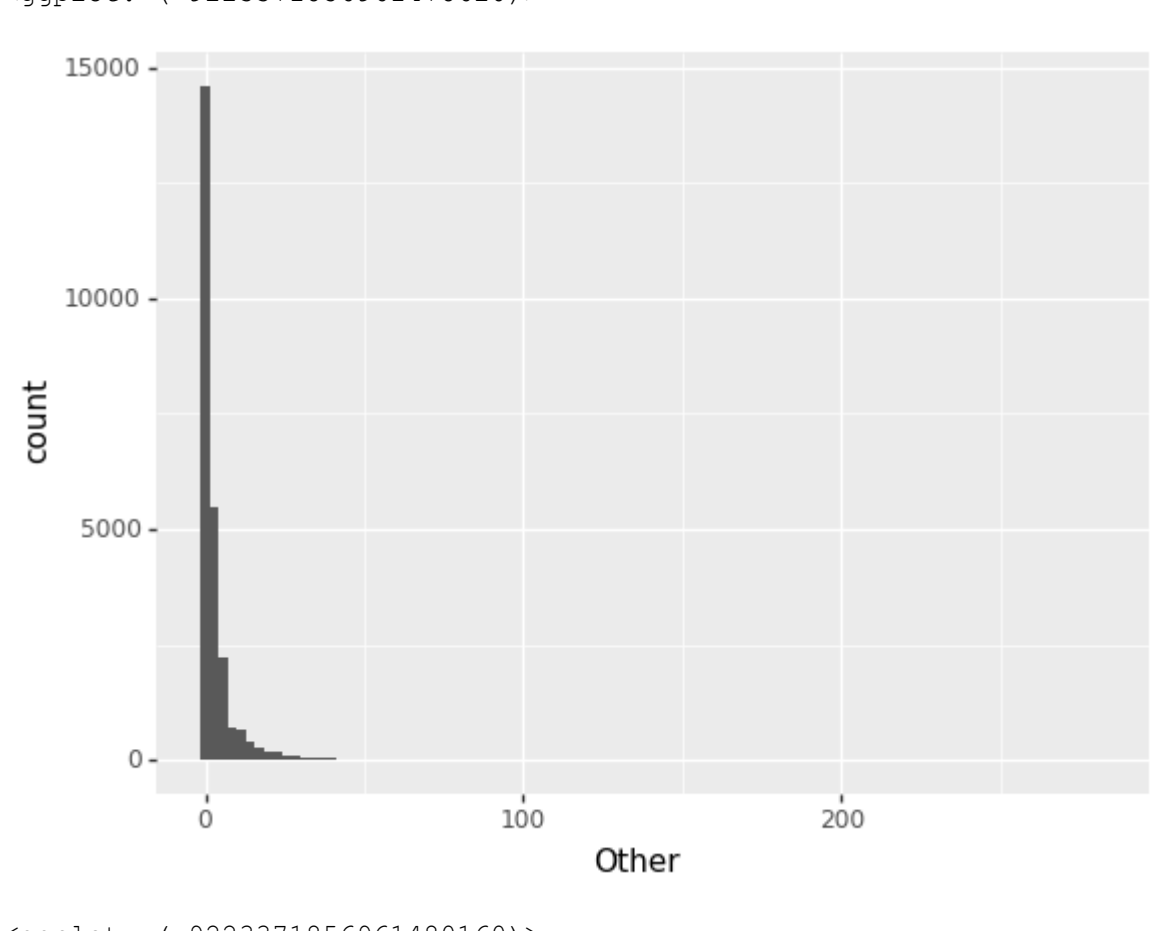
<ggplot: (-9223371856961342288)>



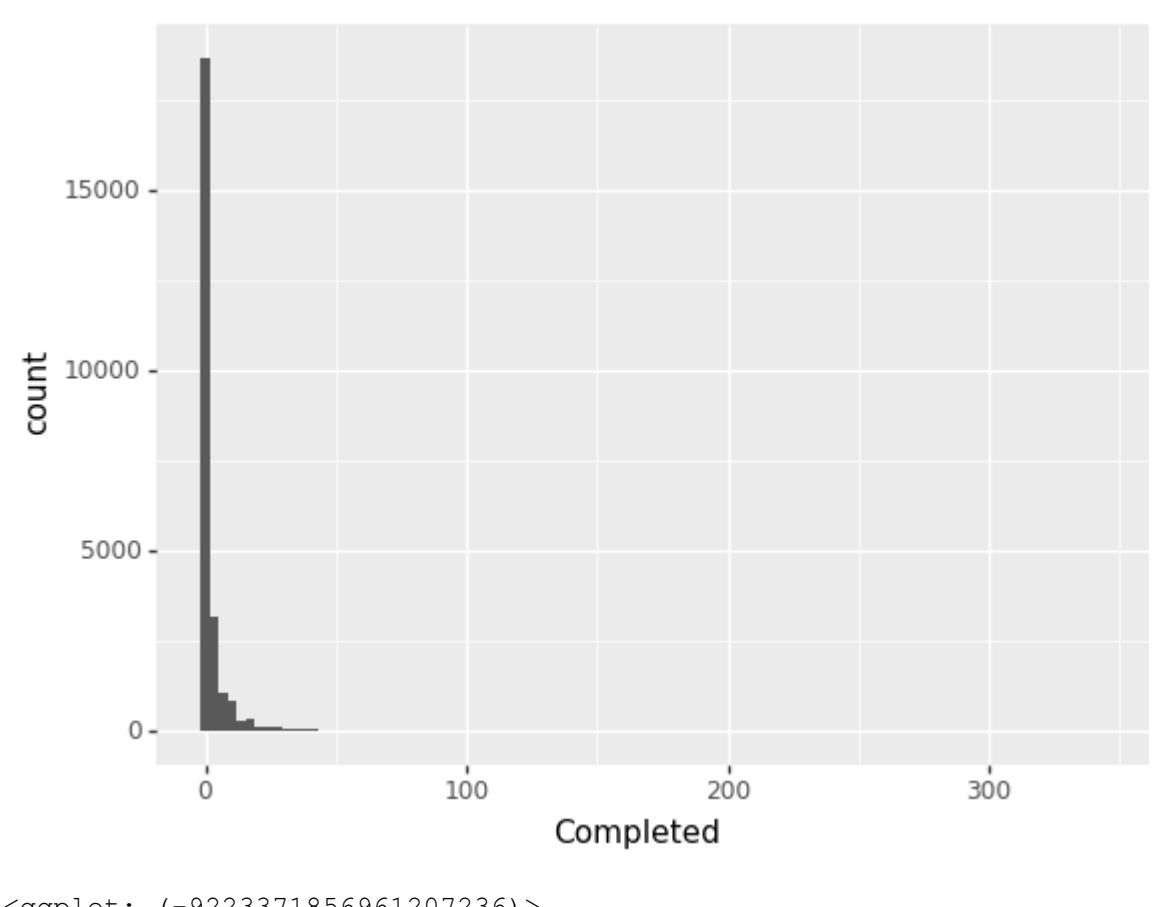
<ggplot: (-9223371856961498716)>



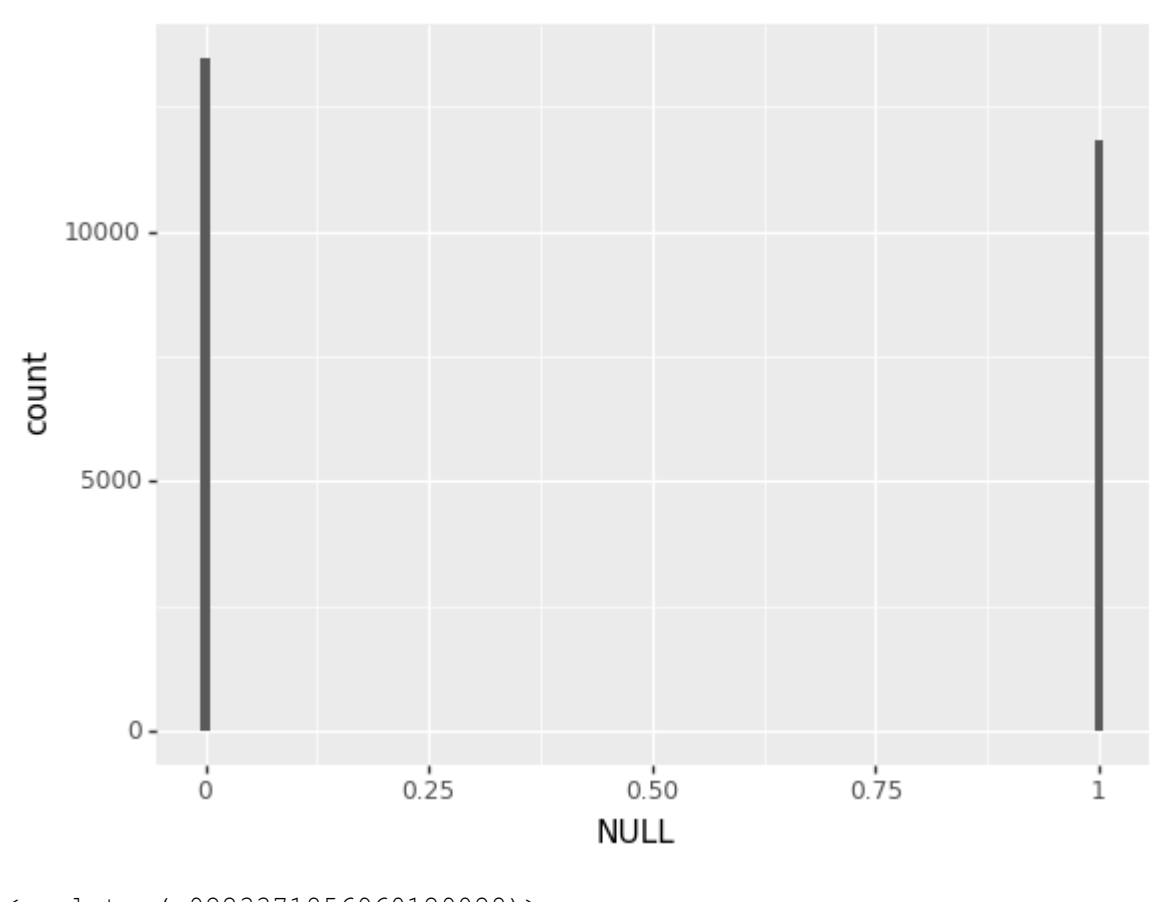
<ggplot: (-9223371856961478620)>



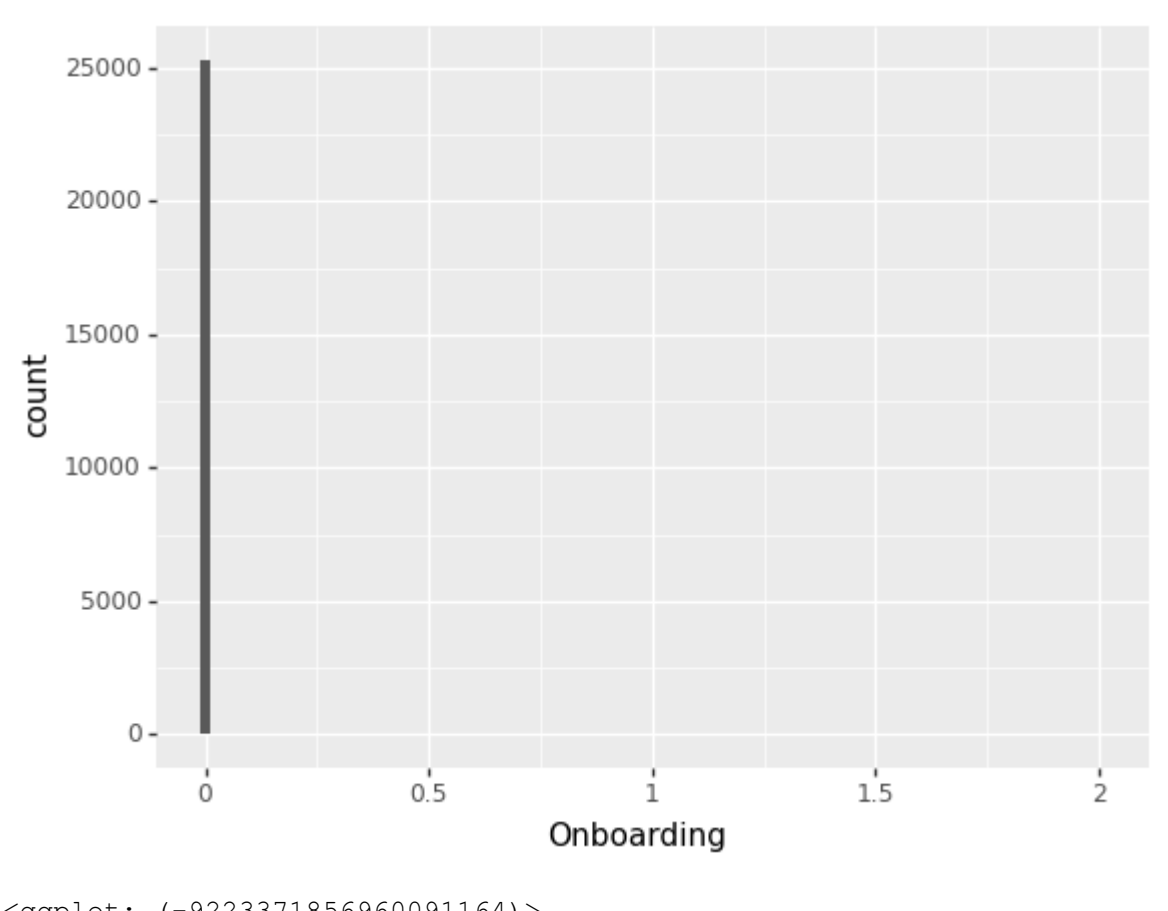
<ggplot: (-9223371856961480160)>



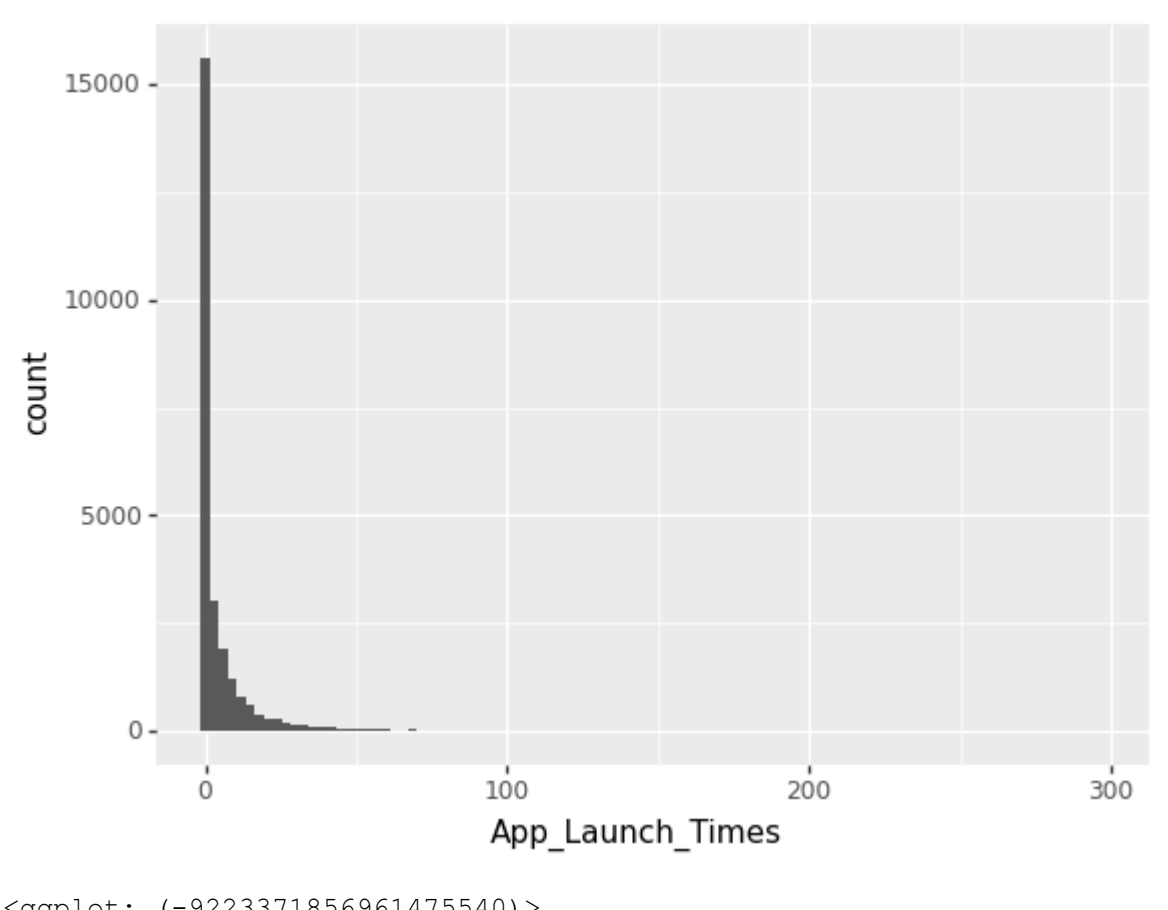
<ggplot: (-9223371856961207236)>



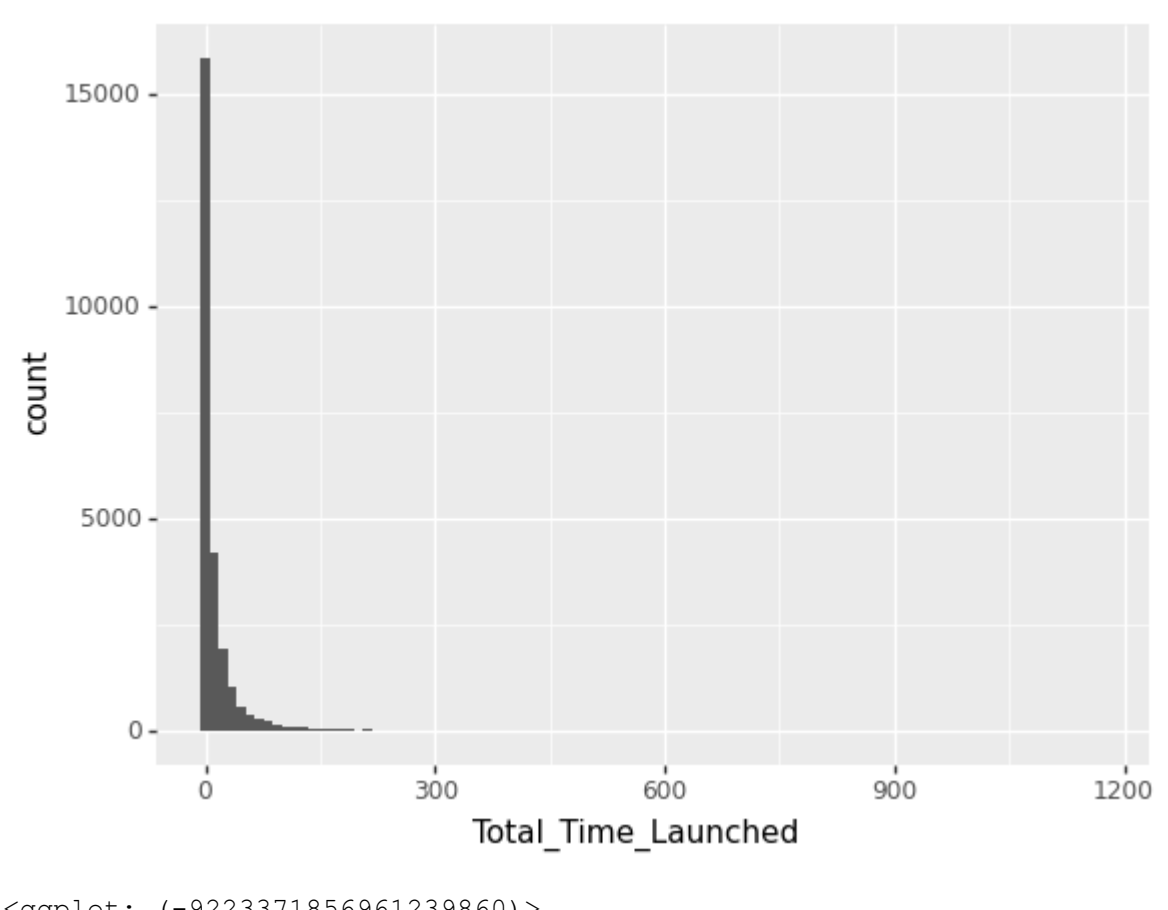
<ggplot: (-9223371856960129028)>



<ggplot: (-9223371856960091164)>



<ggplot: (-9223371856961475540)>



<ggplot: (-9223371856961239860)>

Categorical

- Language
- Subscription_Type
- Subscription_Event_Type
- Purchase_Store
- Demo_User
- Free_Trial_User
- Auto_Renew
- Country
- User_type
- Lead_Platform
- Email_Subscriber
- Push_Notifications

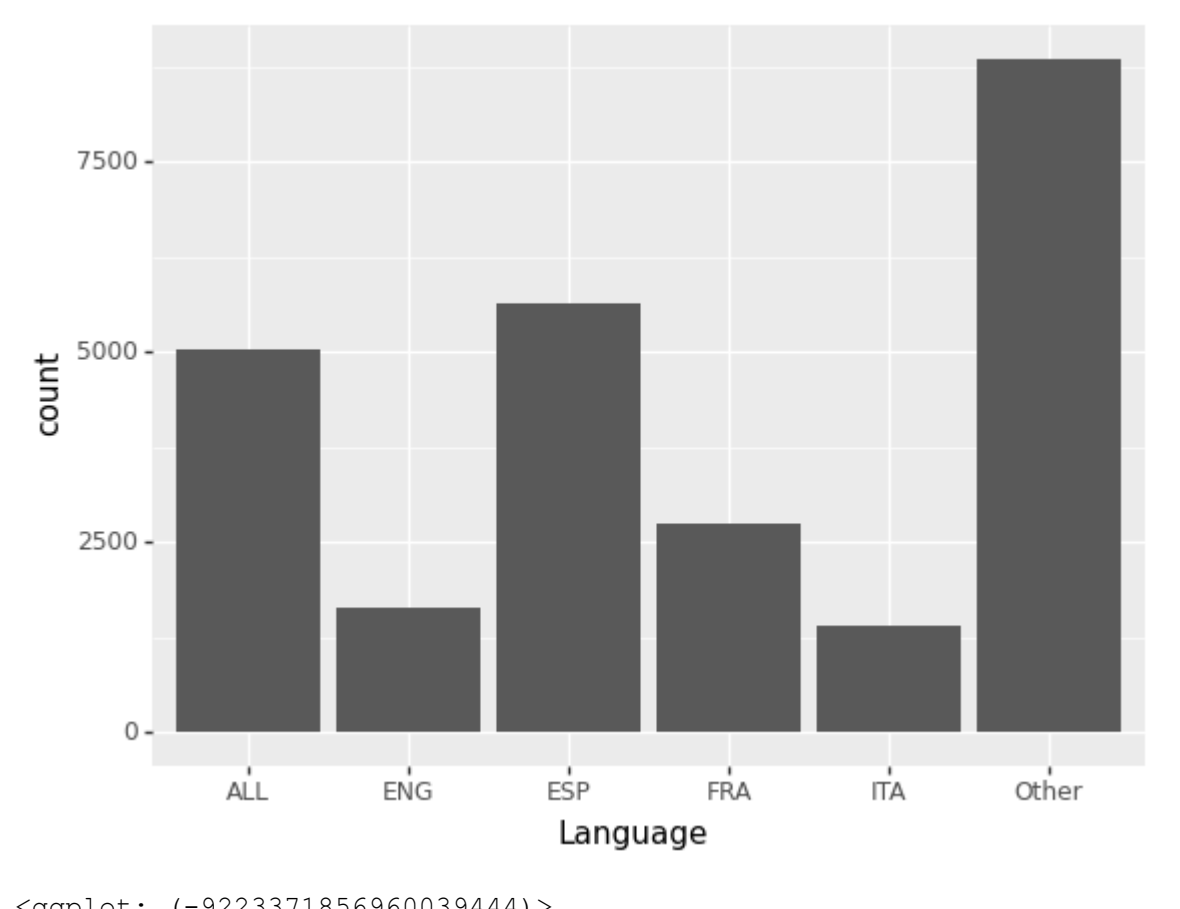
Binary

- Subscription_Type
- Subscription_Event_Type
- Purchase_Store
- Demo_User
- Free_Trial_User
- Auto_Renew
- User_Type
- Email_Subscriber
- Push_Notifications

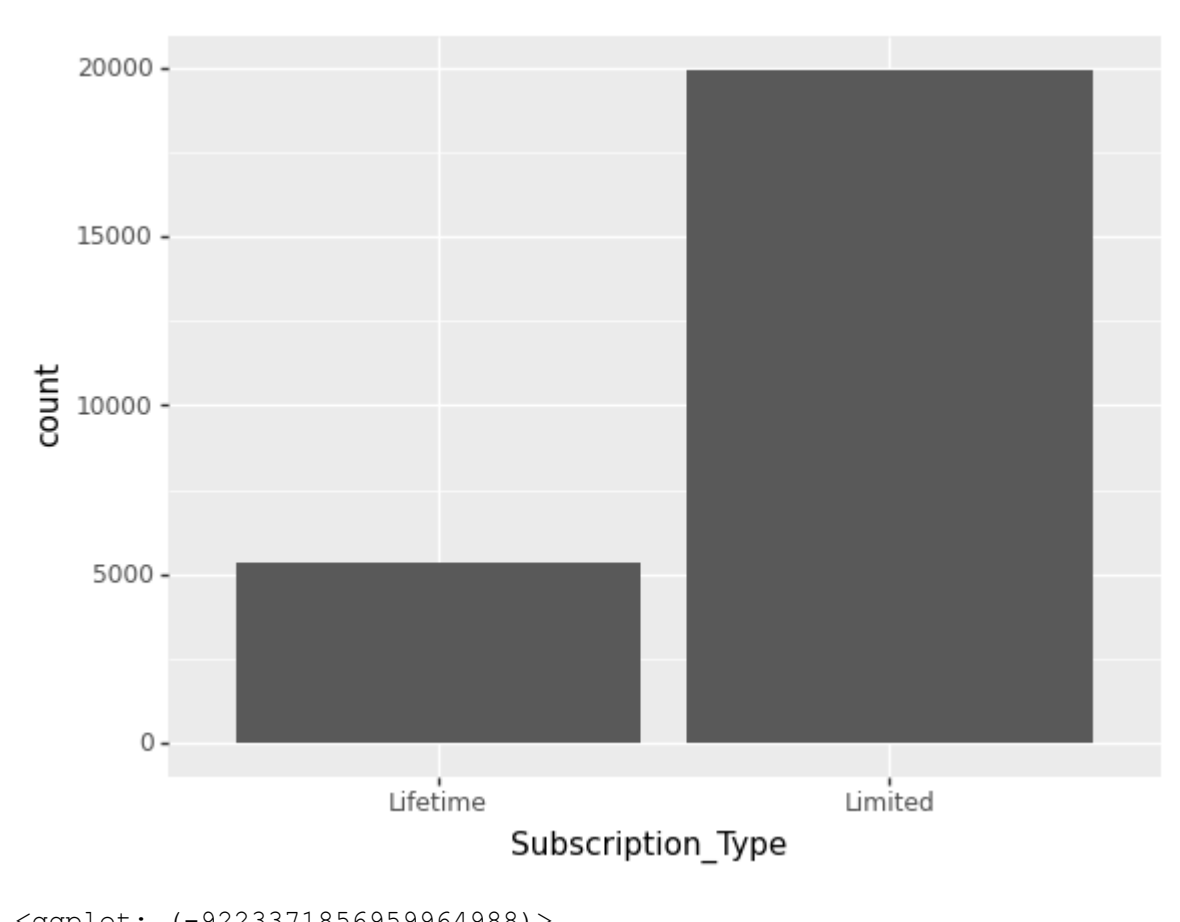
In [7]:
```python
categorical_columns = ['Language', 'Subscription_Type', 'Subscription_Event_Type', 'Purchase_Store', 'D
emo_User', 'Free_Trial_User', 'Auto_Renew', 'Country', 'User_Type', 'Lead_Platform', 'Email_Subscriber'
, 'Push_Notifications']
categorical = data[categorical_columns]
```

In [8]:

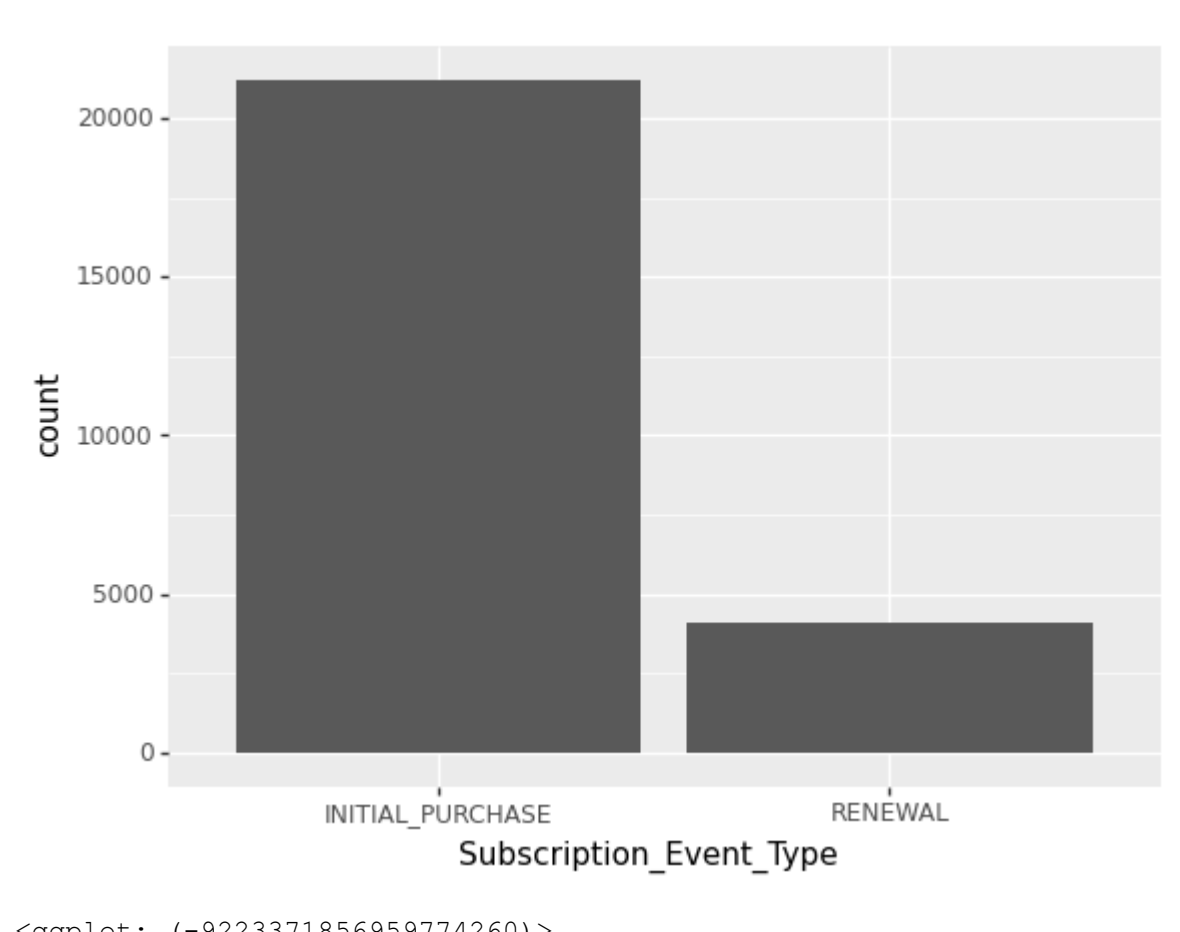
```python
for i in categorical_columns:
    print(ggplot(data, aes(x = i)) + geom_bar())
```



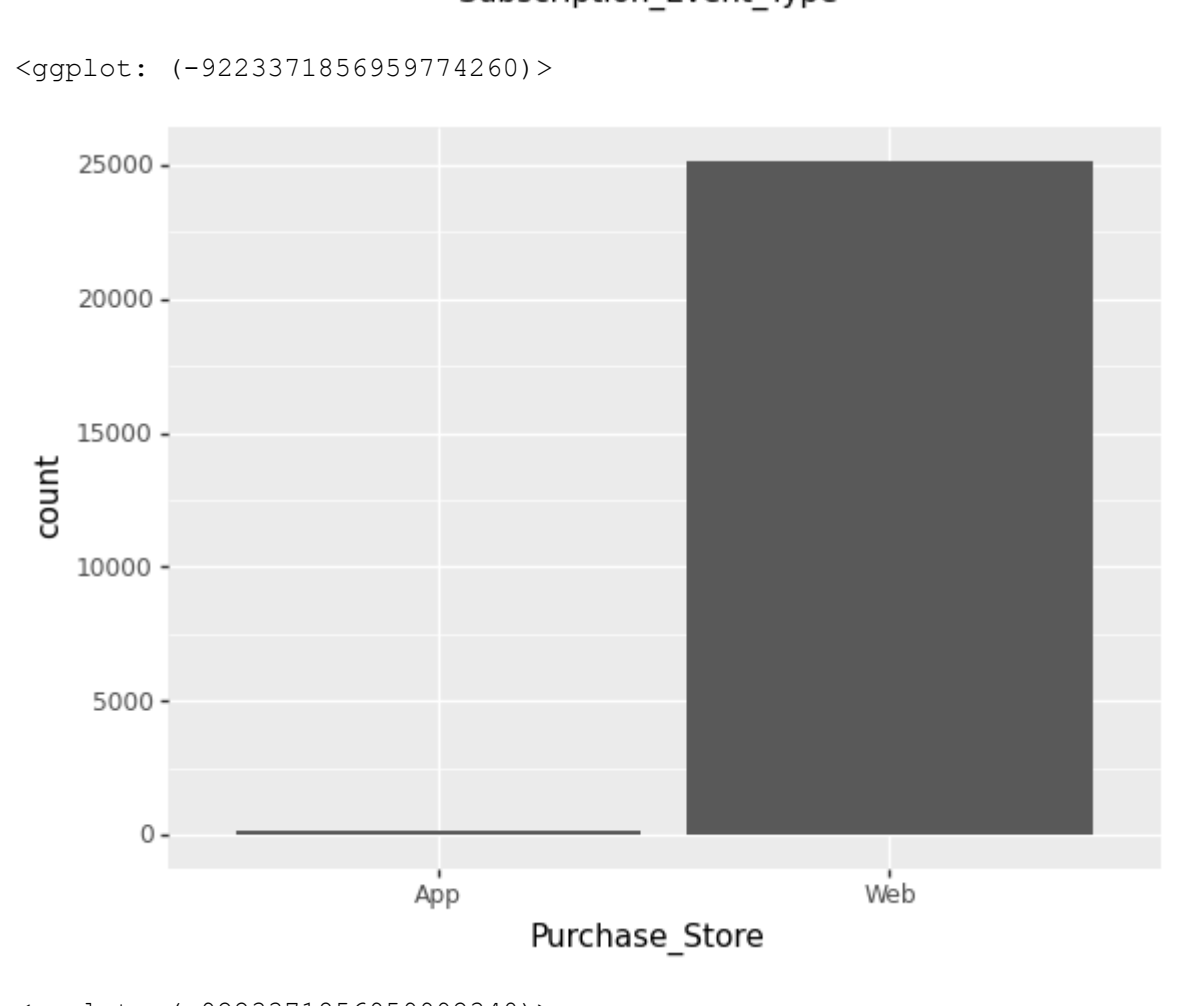

```
<ggplot: (-9223371856960039444)>
```



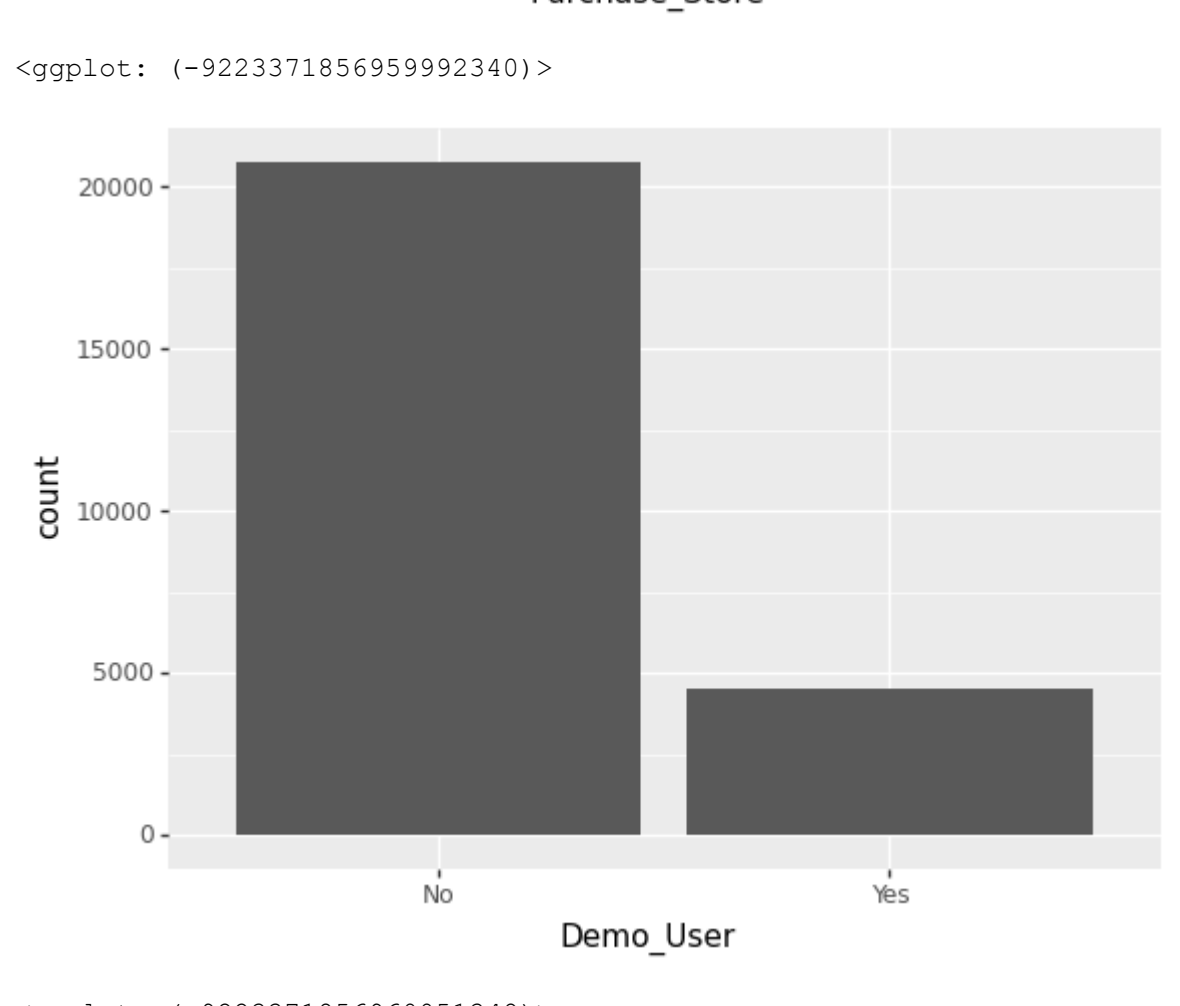

```
<ggplot: (-9223371856959964988)>
```



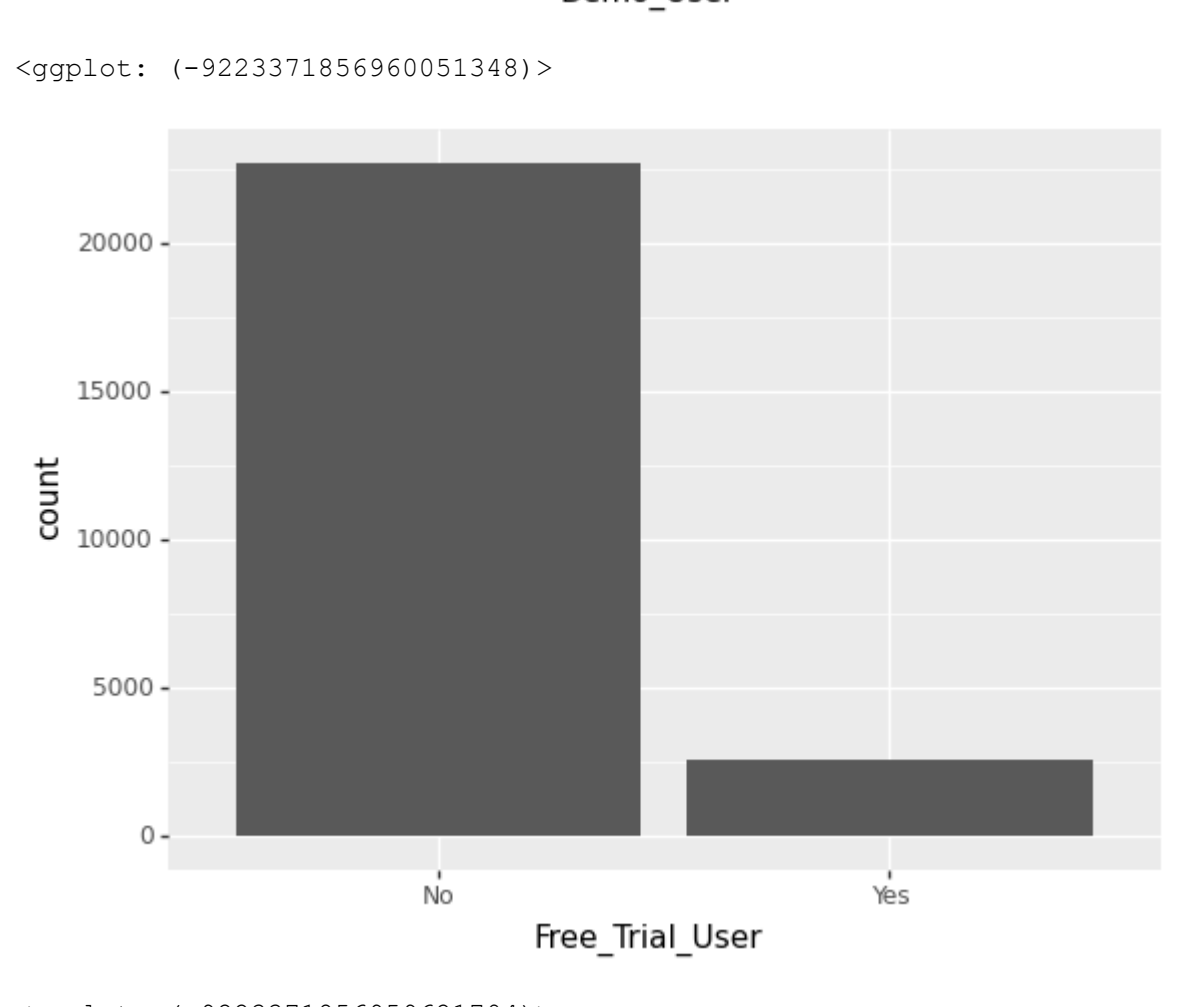

```
<ggplot: (-9223371856959774260)>
```



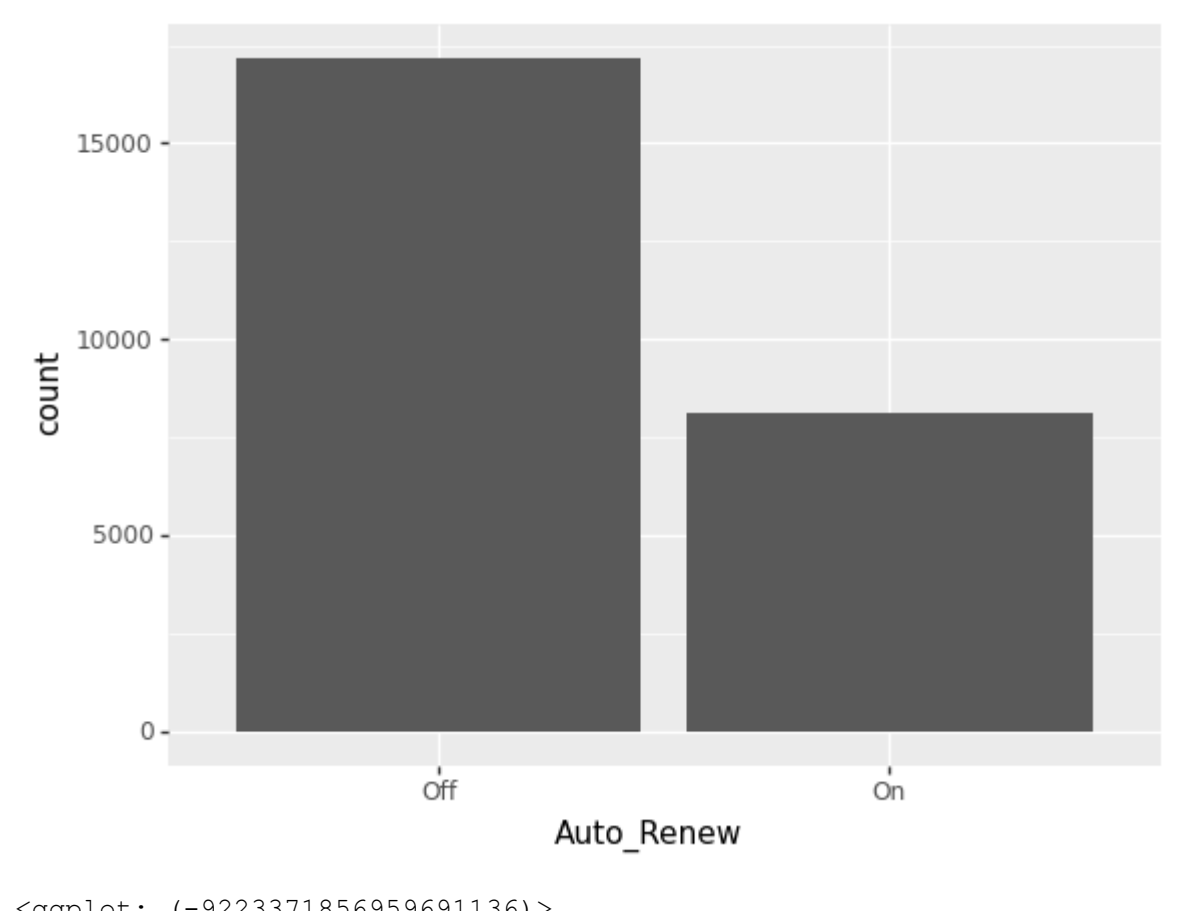

```
<ggplot: (-9223371856959992340)>
```



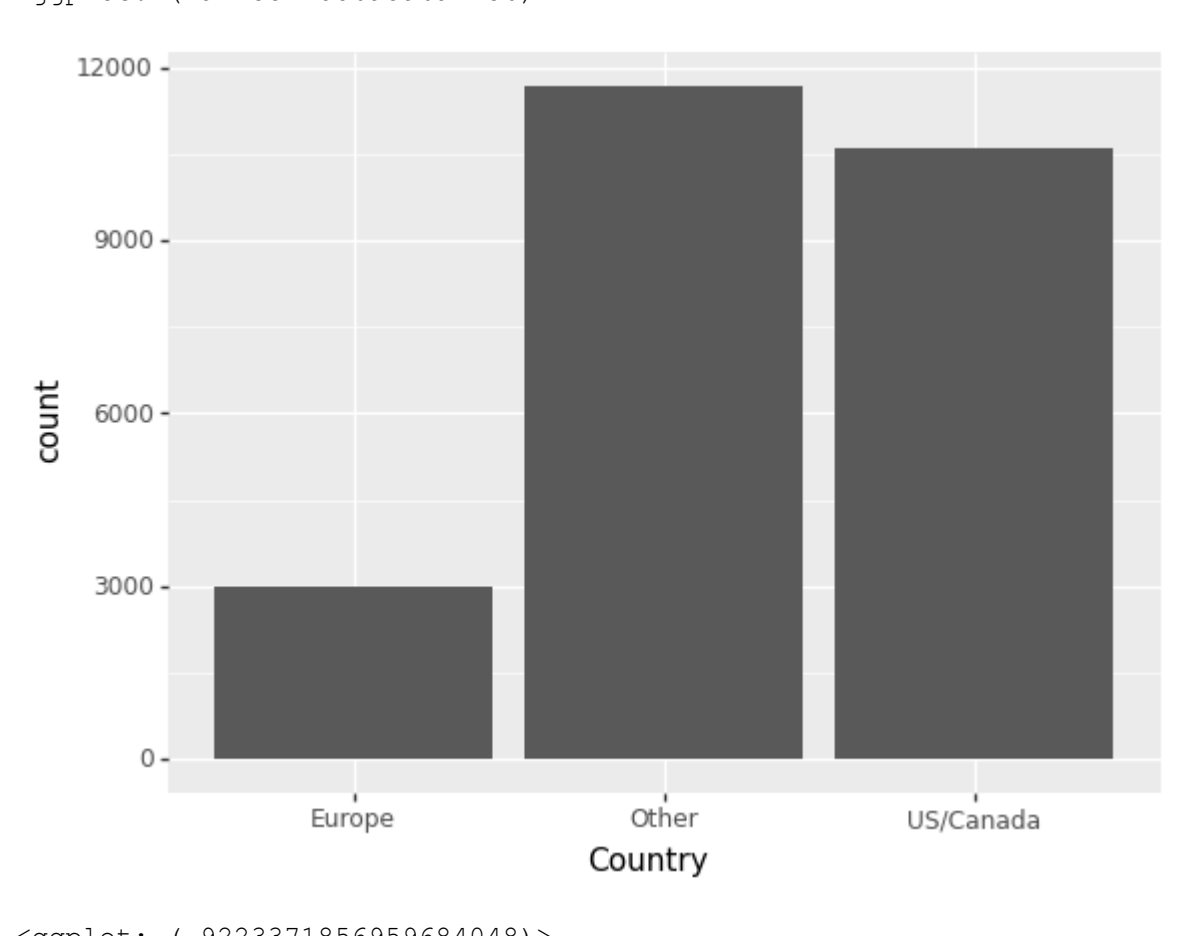

```
<ggplot: (-9223371856960051348)>
```



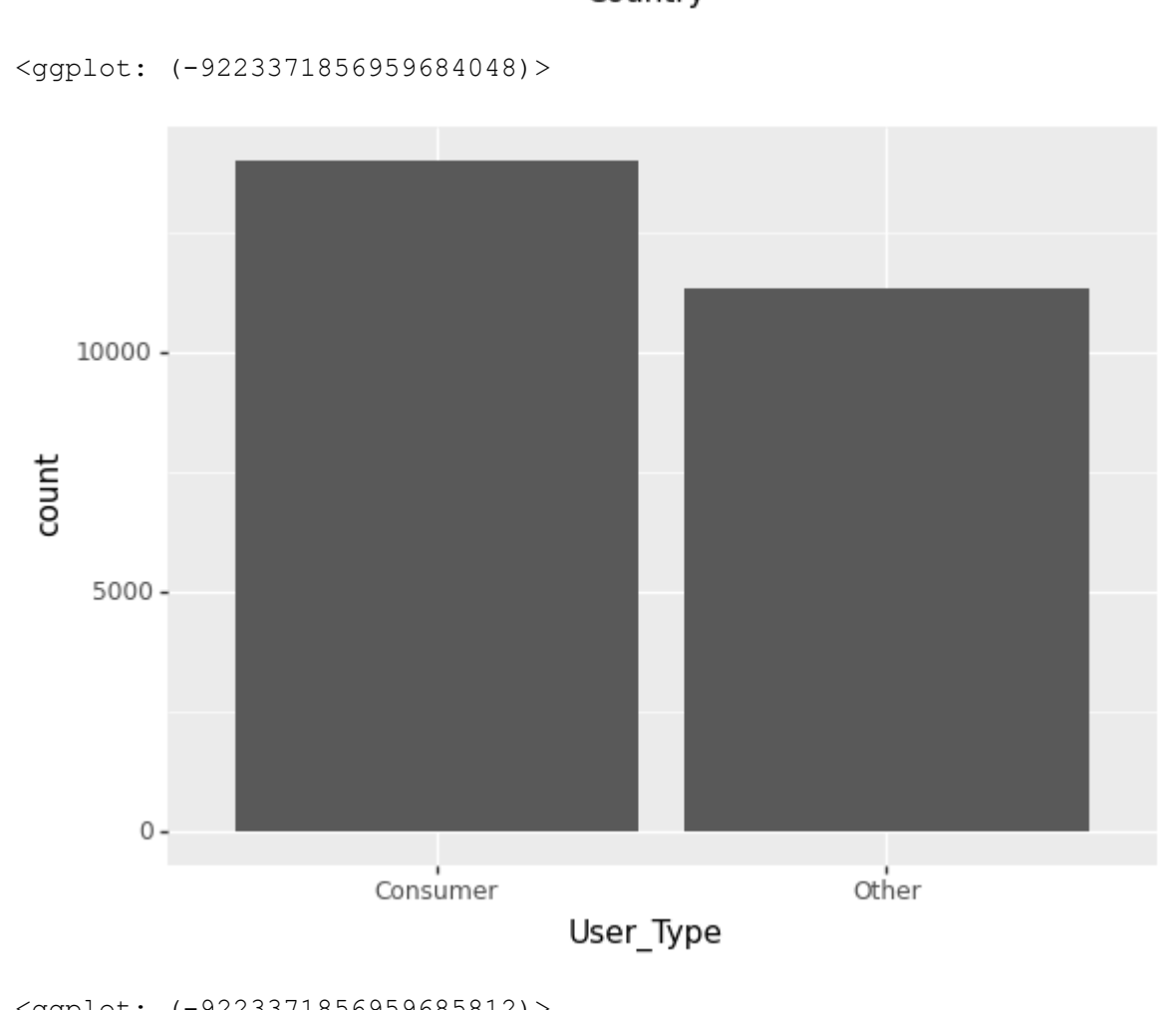

```
<ggplot: (-9223371856959621704)>
```



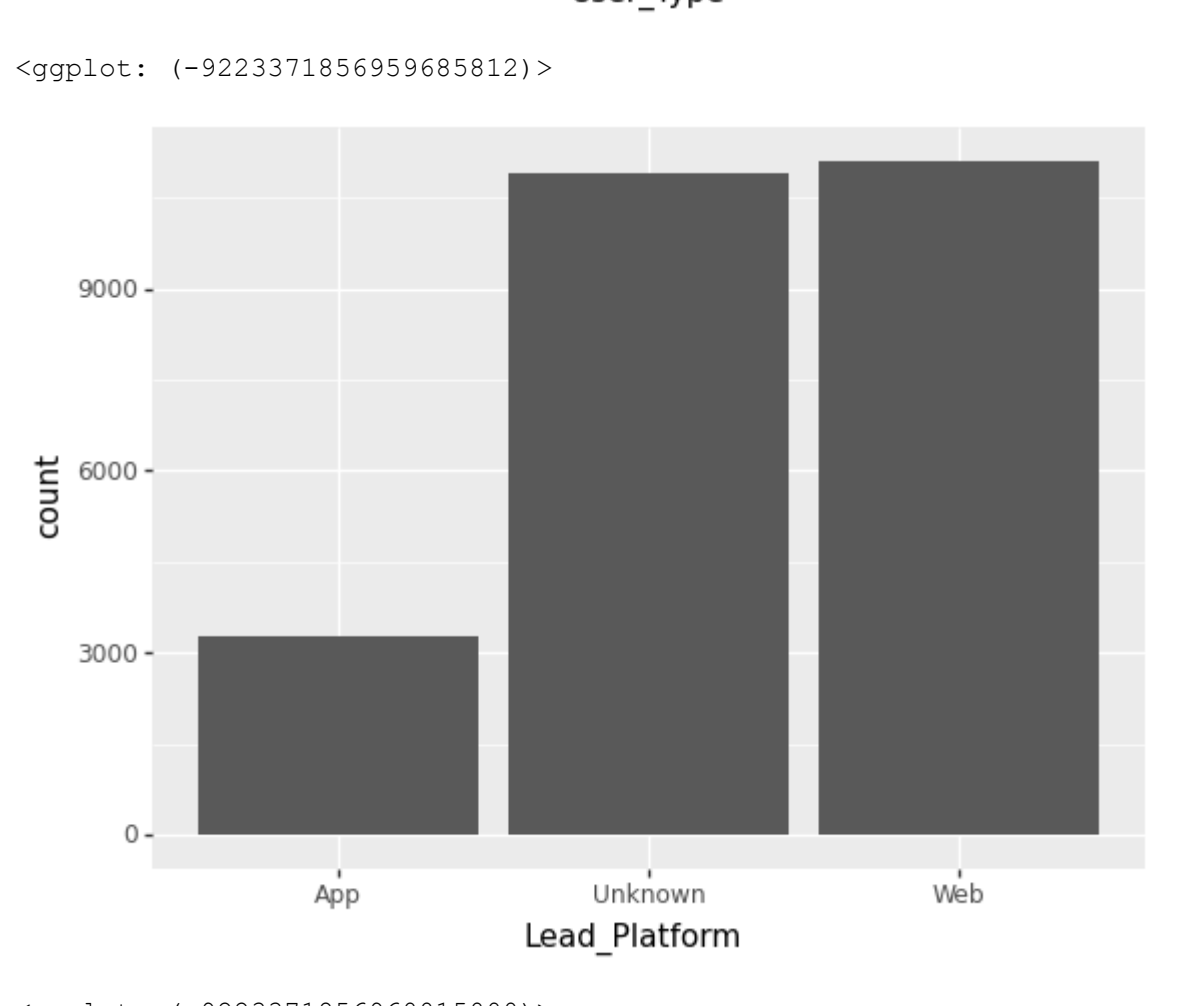

```
<ggplot: (-9223371856959691136)>
```



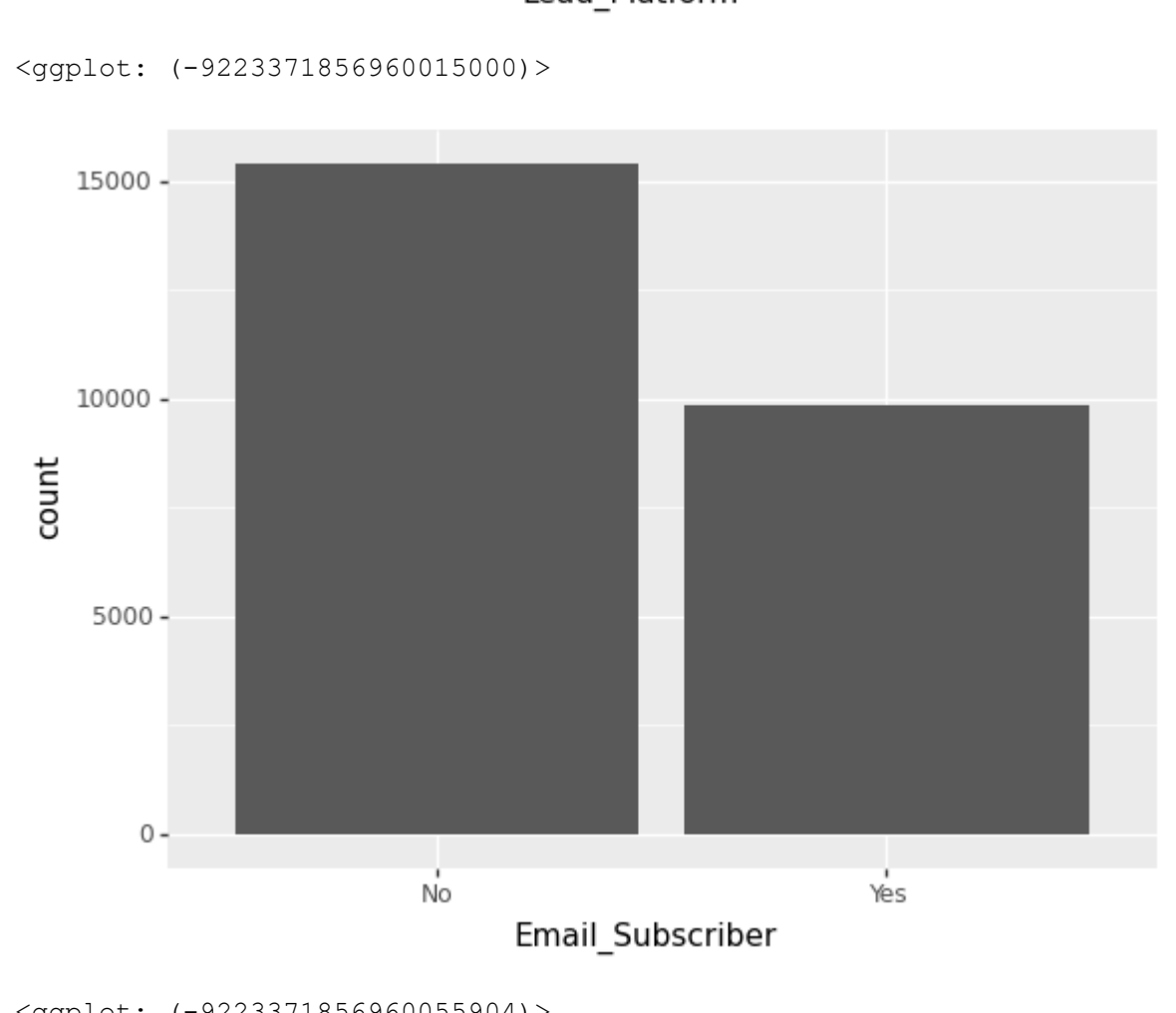

```
<ggplot: (-9223371856959684048)>
```



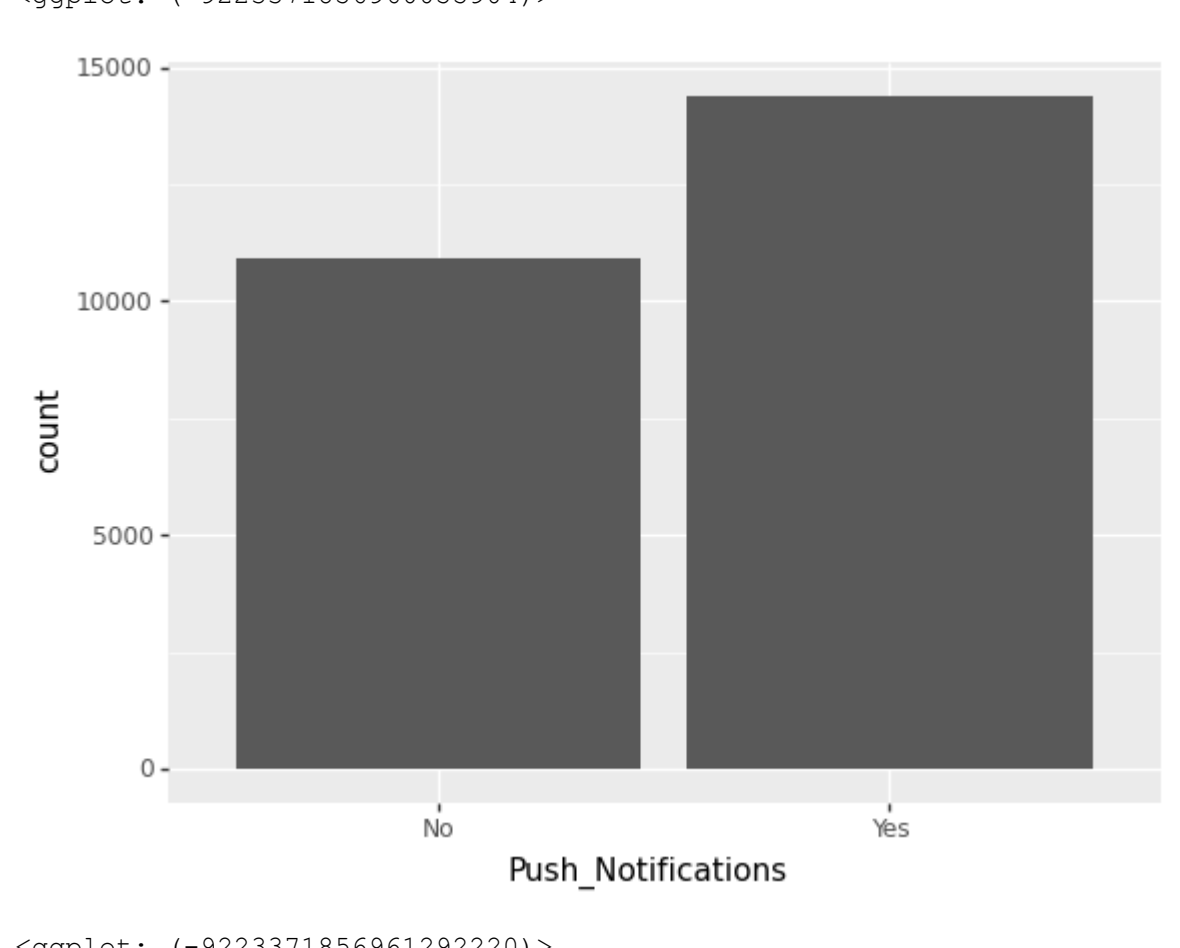

```
<ggplot: (-9223371856959685812)>
```




```
<ggplot: (-9223371856960015000)>
```




```
<ggplot: (-9223371856960055904)>
```




```
<ggplot: (-9223371856961292220)>
```

In [9]:

```python
categorical.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25291 entries, 0 to 25290
Data columns (total 12 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Language                 25291 non-null  object
 1   Subscription_Type        25291 non-null  object
 2   Subscription_Event_Type  25291 non-null  object
 3   Purchase_Store           25291 non-null  object
 4   Demo_User                25291 non-null  object
 5   Free_Trial_User          25291 non-null  object
 6   Auto_Renew               25291 non-null  object
 7   Country                  25291 non-null  object
 8   User_Type                25291 non-null  object
 9   Lead_Platform            25291 non-null  object
 10  Email_Subscriber         25291 non-null  object
 11  Push_Notifications       25291 non-null  object
dtypes: object(12)
memory usage: 2.3+ MB
```

In [10]:

```python
from sklearn.preprocessing import OneHotEncoder

enc = OneHotEncoder()
enc.fit(categorical)
enc_X = enc.transform(categorical).toarray()

enc_df = pd.DataFrame(enc_X, columns = enc.get_feature_names(categorical.columns))
enc_df.head()
```

Out[10]:

	Language_ALL	Language_ENG	Language_ESP	Language_FRA	Language_ITA	Language_Other	Subscription_Type_Lifetime	Subscript
0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
1	0.0	0.0	1.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	1.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	1.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	1.0	0.0	

5 rows × 30 columns

In [11]:

```python
web2 = pd.concat([numerical, enc_df],axis = 1)
web2
```

Out[11]:

	ID	Purchase_Amounts	Duration	Send_Count	Open_Count	Click_Count	Unique_Open_Count	Unique_Click_Count	Start	Oth
0	2	39.00	92	4	3	0	1	0	12	
1	3	0.00	365	1	0	0	0	0	0	
2	6	38.34	92	162	1	0	1	0	3	
3	7	79.00	113	2	0	0	0	0	7	
4	8	38.40	92	25	17	4	7	2	13	
...	...	...	...	...	...	...	...	...	...	
25286	39993	19.63	31	40	0	0	0	0	0	
25287	39994	212.13	28781	6	2	0	2	0	0	
25288	39995	0.00	92	0	0	0	0	0	0	
25289	39996	48.36	299	1	0	0	0	0	0	
25290	39998	12.40	93	0	0	0	0	0	0	

25291 rows × 45 columns

PCA

Web

In [12]:

```python
features = web2
features
```

Out[12]:

	ID	Purchase_Amounts	Duration	Send_Count	Open_Count	Click_Count	Unique_Open_Count	Unique_Click_Count	Start	Oth
0	2	39.00	92	4	3	0	1	0	12	
1	3	0.00	365	1	0	0	0	0	0	
2	6	38.34	92	162	1	0	1	0	3	
3	7	79.00	113	2	0	0	0	0	7	
4	8	38.40	92	25	17	4	7	2	13	
...	...	...	...	...	...	...	...	...	...	
25286	39993	19.63	31	40	0	0	0	0	0	
25287	39994	212.13	28781	6	2	0	2	0	0	
25288	39995	0.00	92	0	0	0	0	0	0	
25289	39996	48.36	299	1	0	0	0	0	0	
25290	39998	12.40	93	0	0	0	0	0	0	

25291 rows × 45 columns

In [13]:

```python
del features['ID']
```

In [14]:

```python
features.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25291 entries, 0 to 25290
Data columns (total 44 columns):
 #   Column                                   Non-Null Count  Dtype
---  ------                                   --------------  -----
 0   Purchase_Amounts                         25291 non-null  float64
 1   Duration                                 25291 non-null  int64
 2   Send_Count                               25291 non-null  int64
 3   Open_Count                               25291 non-null  int64
 4   Click_Count                              25291 non-null  int64
 5   Unique_Open_Count                        25291 non-null  int64
 6   Unique_Click_Count                       25291 non-null  int64
 7   Start                                    25291 non-null  int64
 8   Other                                    25291 non-null  int64
 9   Completed                                25291 non-null  int64
 10  NULL                                     25291 non-null  int64
 11  Onboarding                               25291 non-null  int64
 12  App_Launch_Times                         25291 non-null  int64
 13  Total_Time_Launched                      25291 non-null  int64
 14  Language_ALL                             25291 non-null  float64
 15  Language_ENG                             25291 non-null  float64
 16  Language_ESP                             25291 non-null  float64
 17  Language_FRA                             25291 non-null  float64
 18  Language_ITA                             25291 non-null  float64
 19  Language_Other                           25291 non-null  float64
 20  Subscription_Type_Lifetime               25291 non-null  float64
 21  Subscription_Type_Limited                25291 non-null  float64
 22  Subscription_Event_Type_INITIAL_PURCHASE 25291 non-null  float64
 23  Subscription_Event_Type_RENEWAL          25291 non-null  float64
 24  Purchase_Store_App                       25291 non-null  float64
 25  Purchase_Store_Web                       25291 non-null  float64
 26  Demo_User_No                             25291 non-null  float64
 27  Demo_User_Yes                            25291 non-null  float64
 28  Free_Trial_User_No                       25291 non-null  float64
 29  Free_Trial_User_Yes                      25291 non-null  float64
 30  Auto_Renew_Off                           25291 non-null  float64
 31  Auto_Renew_On                            25291 non-null  float64
 32  Country_Europe                           25291 non-null  float64
 33  Country_Other                            25291 non-null  float64
 34  Country_US/Canada                        25291 non-null  float64
 35  User_Type_Consumer                       25291 non-null  float64
 36  User_Type_Other                          25291 non-null  float64
 37  Lead_Platform_App                        25291 non-null  float64
 38  Lead_Platform_Unknown                    25291 non-null  float64
 39  Lead_Platform_Web                        25291 non-null  float64
 40  Email_Subscriber_No                      25291 non-null  float64
 41  Email_Subscriber_Yes                     25291 non-null  float64
 42  Push_Notifications_No                    25291 non-null  float64
 43  Push_Notifications_Yes                   25291 non-null  float64
dtypes: float64(31), int64(13)
memory usage: 8.5 MB
```

In [15]:

```python
from sklearn.preprocessing import StandardScaler
features = StandardScaler().fit_transform(features)
features
```

Out[15]:

```
array([[-0.40008706, -0.53303425, -0.47392336, ..., -0.80082053,
        -0.87043788,  0.87043788],
       [-0.90130578, -0.50964295, -0.53725744, ...,  1.24871924,
        -0.87043788,  0.87043788],
       [-0.40856922, -0.53303425,  2.86167149, ...,  1.24871924,
        -0.87043788,  0.87043788],
       ...,
       [-0.90130578, -0.53303425, -0.5583688 , ..., -0.80082053,
         1.14884706, -1.14884706],
       [-0.27979457, -0.51529799, -0.53725744, ..., -0.80082053,
         1.14884706, -1.14884706],
       [-0.74194393, -0.53294857, -0.5583688 , ..., -0.80082053,
        -0.87043788,  0.87043788]])
```

In [16]:

```python
from sklearn.preprocessing import Normalizer
features = Normalizer().fit_transform(features)

features
```

Out[16]:

```
array([[-0.07449584, -0.09925049, -0.08824409, ..., -0.14911205,
        -0.16207473,  0.16207473],
       [-0.1415996 , -0.08006743, -0.0844058 , ...,  0.19617998,
        -0.1367501 ,  0.1367501 ],
       [-0.05908559, -0.07708521,  0.41384312, ...,  0.18058462,
        -0.12587913,  0.12587913],
       ...,
       [-0.17617471, -0.10419011, -0.10914216, ..., -0.15653325,
         0.22456063, -0.22456063],
       [-0.06482256, -0.11938378, -0.12447132, ..., -0.18553338,
         0.26616386, -0.26616386],
       [-0.1278343 , -0.09182514, -0.09620496, ..., -0.13797853,
        -0.14997335,  0.14997335]])
```

In [17]:

```python
from sklearn.decomposition import PCA #reducing the variables to 2
pca = PCA(n_components=2)
principalComponents = pca.fit_transform(features)

principalComponents
```

Out[17]:

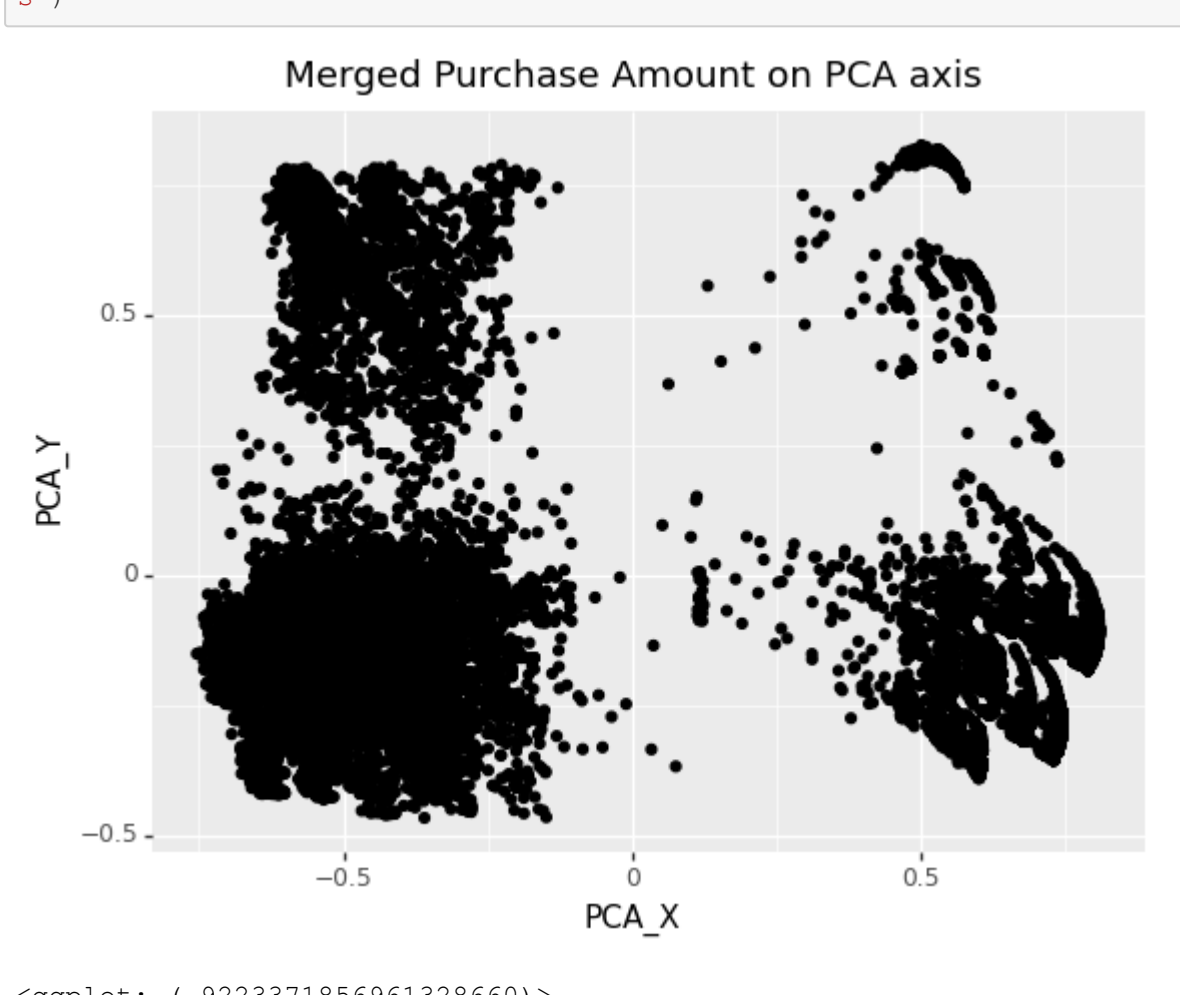
```
array([[-0.29219426, -0.1609633 ],
       [-0.58199422, -0.13976677],
       [-0.62054248, -0.01514089],
       ...,
       [ 0.65562786, -0.11196975],
       [ 0.80576289, -0.13203789],
       [-0.29213446, -0.3224244 ]])
```

In [18]:

```python
web2['PCA_X'] = principalComponents[:,0]
web2['PCA_Y'] = principalComponents[:,1]
```

In [19]:

```python
ggplot(web2, aes(x = 'PCA_X', y = 'PCA_Y')) + geom_point() + ggtitle('Merged Purchase Amount on PCA axis')
```



Out[19]:

```
<ggplot: (-9223371856961328660)>
```

In [20]:

```python
pca.explained_variance_ratio_
#only explaines about 45% of data
```

Out[20]:

```
array([0.32865293, 0.1352765 ])
```

In [21]:

```python
pca = PCA(.90)
```

In [22]:

```python
principalComponents = pca.fit_transform(features)
```

In [23]:

```python
pca.explained_variance_ratio_
```

Out[23]:

```
array([0.32865293, 0.1352765 , 0.06827549, 0.06314286, 0.05766444,
       0.04430393, 0.04119663, 0.03579236, 0.03456028, 0.02930314,
       0.027936  , 0.02601676, 0.02462589])
```

In [24]:

```python
for i in range(0,13):
    name = 'PCA' + str(i+1)

    web2[name] = principalComponents[:,i]
```

In [25]:

```python
web2 = pd.concat([web2, categorical], axis = 1)
```

In [26]:

```python
web2
```

Out[26]:

	Purchase_Amounts	Duration	Send_Count	Open_Count	Click_Count	Unique_Open_Count	Unique_Click_Count	Start	Other	Con
0	39.00	92	4	3	0	1	0	12	25	
1	0.00	365	1	0	0	0	0	0	2	
2	38.34	92	162	1	0	1	0	3	21	
3	79.00	113	2	0	0	0	0	7	9	
4	38.40	92	25	17	4	7	2	13	21	
...	...	...	...	...	...	...	...	...	...	
25286	19.63	31	40	0	0	0	0	0	0	
25287	212.13	28781	6	2	0	2	0	0	0	
25288	0.00	92	0	0	0	0	0	0	0	
25289	48.36	299	1	0	0	0	0	0	0	
25290	12.40	93	0	0	0	0	0	0	1	

25291 rows × 71 columns

In [27]:

```python
web2.to_csv('Subscriber Information (Clean) Version 5.csv', index = False)
```