**Introduction**The purpose of this analytics exercise is to isolate IP addresses belonging to VPN apps in order to improve accuracy of household identification model.   
  
VPN stands for “Virtual Private Network”. A VPN server hides the user’s IP address and online activity by letting the network redirect it through a specially configured remote server run by a VPN host. Apps that provide this service are referred to as “VPN apps”.

VPN apps mask the true location of the user, thereby reducing the integrity of the training data used for the model. This documentation details how we can attempt to identify and exclude IP addresses which are associated with VPN apps.   
  
Our primary source of app data comes from this path: 's3a://ada-prod-data/etl/data/ref/bundle/segment/monthly/all/all/'

**Data Description**

Here’s the schema for the app data we use for the primary analysis:

# |-- app\_name: string (nullable = true)

# |-- bundle: string (nullable = true)

# |-- description: string (nullable = true)

# |-- genres: array (nullable = true)

# | |-- element: string (containsNull = true)

# |-- os\_platform: string (nullable = true)

# |-- segment: array (nullable = true)

# | |-- element: string (containsNull = true)

# |-- genre\_segment: array (nullable = true)

# | |-- element: struct (containsNull = true)

# | | |-- genres: string (nullable = true)

# | | |-- segment: string (nullable = true)

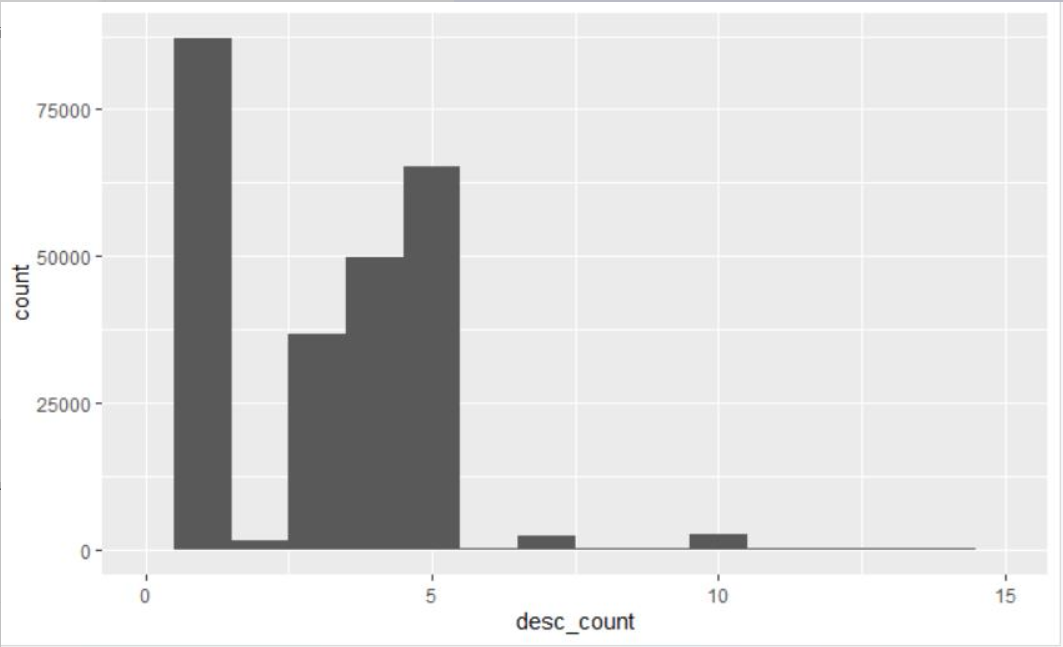
For this analysis, we only need three columns: bundle, app\_name, description.   
  
**“bundle”** column is a unique identifier for each app, which could be a numerical string or a string of alphabets depending on the phone’s OS.   
  
**“app\_name”** column is a string containing the name of the app.   
  
**“description”** column is a string containing detailed description of the app function. This is usually the promotional description text displayed under the app at both Android and Apple app stores.   
  
We add two more column, **“app\_count”** and **“desc\_count”** to count how many times the string “vpn” occurs in app name and app description respectively.   
  
Here is the general distribution of the occurrence frequency in app description and app name:  
  


Figure 1: A distribution of how many times the string "vpn" occurs in app description

From figure 1, we can see that the most common frequency is a single time, which intuitively makes sense. However, this is not conclusive proof of the app being an VPN app. Often times, it is a disclaimer such as: *“This app will not work with VPNs”*  
  
However, for occurrence frequency of 3 and above, we are able to identify apps whose primary function is to provide VPN services.

As for the app\_count distribution, single occurrence of the string “vpn” is most common, which intuitively makes sense because it would be unusual for an app name to contain the string “vpn” more than one time. We can see the distribution in figure 2.

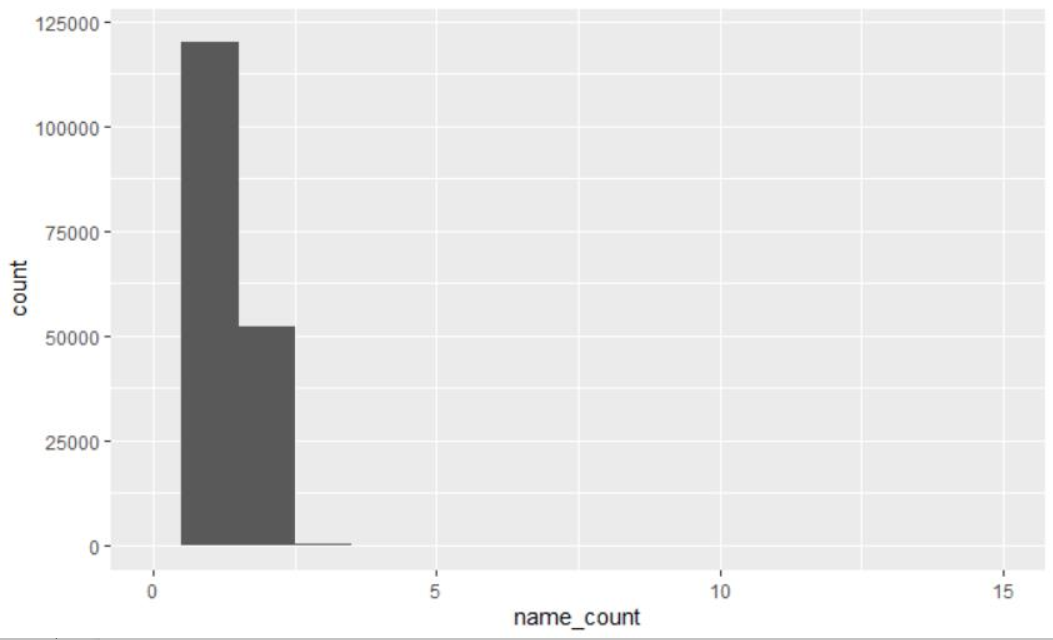


Figure 2: Frequency distribution of the string "vpn" in app name

**VPN IP Identification Methodology**

Since there is no pre-labelled “ground truth” training data, we have to manually read the description and confirm whether an app is a VPN app or not. Filtering apps by name only results in 28% less IP addresses because a VPN app can be named *“Avast Anti-Virus”* or *“Proxy Browser”* or *“Website Unblocker”* etc.   
  
Filtering by single occurrence of the string *“vpn”* in the app description is also not desirable because it yields false positives.   
  
Therefore, the proposed methodology for best outcome is:  
  
Step 1: Initially filter apps that contain the string *“vpn”* in name or app description *at least once.*

Step 2: Filter further using the *“desc\_count”* column with a threshold of 3 or greater occurrence frequency.

*Figure 3: Total Count of IP addresses by country*

Finally, would encourage increasing the *“desc\_count”* threshold to test if it yields better results for the household model, if not, revert back to using the originally propose threshold of 3 or greater.