Scenario: You are contracted by a Westeros software developer with a Health and Wellness app looking to break into the Essos market. Based on the provided data set your objective is to advise the client on the following:

1. Who are the current market leaders in the Health and Wellness category and why are they successful? HINT: what metrics support your findings?
2. What market segments should your client target? The client is interested in opportunities presented in targeting based on location, current app usage behaviors, phone brand, and other demographic information available. HINT: do not limit yourself to counts and histograms.
3. The client is interested in partnering with apps outside the Health and Wellness space to drive cross promotion. Where are the biggest opportunities? Please create some short bullet points to support the recommendation in both data science and business contexts.

1. Who are the current market leaders in the Health and Wellness category and why are they successful? HINT: what metrics support your findings?

Logic of thinking: I define current market leaders in Health and Wellness Category as the ones have the most active unique users. That's currently they have the largest market share in the Health and Wellness Category.

I assume being successful in Health and Wellness category as having the most market share. The below 10 app_id have the most active unique users. So I think they are successful and are the current market leaders.

The metrics I use to determine market leaders are:

1. They have the most active usage (i.e is_active is the most).
2. They have the most unique users, that's the unique devices that have the app installed and active.

Result: The following 10 app_ids are the current market leaders:

```
In [13]: # Show the current top 10 market leaders in Health and Wellness
         run_query(sql).head(10)
```

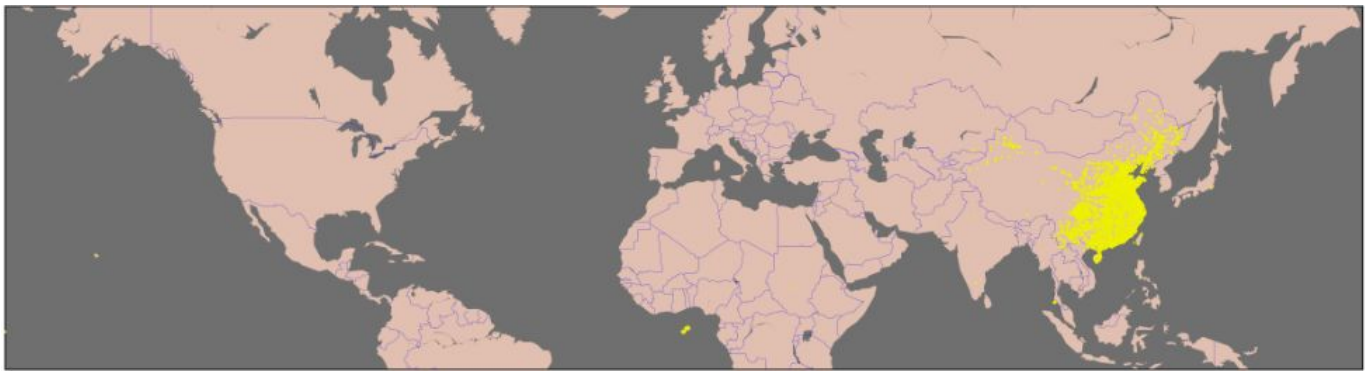
Out[13]:

	app_category	app_id	sum(is_active)	count(distinct device_id)
0	Health and Wellness	1088227414300337900	39943	5356
1	Health and Wellness	3683147815759994238	42951	4687
2	Health and Wellness	6868430133575209713	33814	2686
3	Health and Wellness	-1633887856876571208	31330	2243
4	Health and Wellness	-974457023668610292	23821	1743
5	Health and Wellness	5909705644131785817	27706	1568
6	Health and Wellness	6666573790957269996	13107	1119
7	Health and Wellness	-653184325010919369	5482	1067
8	Health and Wellness	-4986139885405704	25904	980
9	Health and Wellness	-3507529970483852351	9762	965

2. What market segments should your client target? The client is interested in opportunities presented in targeting based on location, current app usage behaviors, phone brand, and other demographic information available. HINT: do not limit yourself to counts and histograms.

Location: For users have Health and Wellness app installed, where are they located?

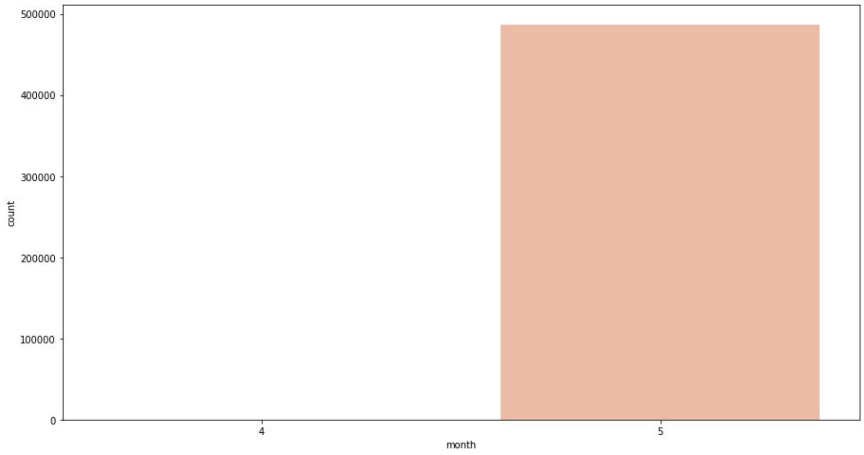
- Checking the Health and Wellness Apps unique user usage by location below, we saw that China has the most active unique users in using Health and Wellness apps (highlighted yellow below).
- There are some dots in the middle of the map, with longitude = 0 and latitude = 0. I think those are the users who don't want to share their location. Based on the volume, it's very small.
- Still conclude that China has the most active unique users in using Health and Wellness app.



App Usage Behaviors: For users have Health and Wellness app installed, which month they use more often? Which day they use more often? What time they use more often?

Logic of thinking: Merged event and app_event_category dataset together to get the device usage information for Health and Wellness app category. Noticed that timestamp in the event dataset is a string. I converted it to datetime so that I can get the week of day and hour when uses used the app.

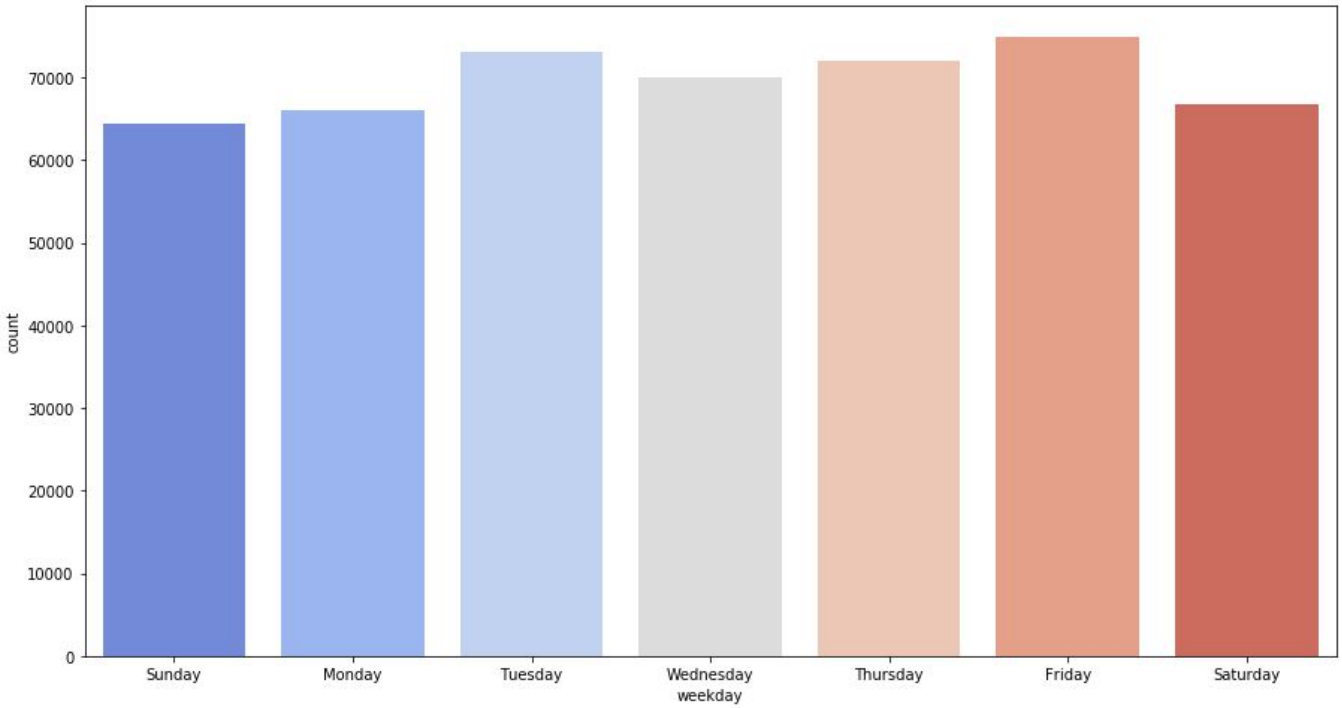
- Check the month usage: Count the number of events for each month for Health and Wellness app:




```
In [105]: device.month.value_counts()

Out[105]: 5    487132
          4       115
          Name: month, dtype: int64
```

Actually found out that the data only mostly have May data. So we can't get usage behavior by month.

- Check the day usage: Count the number of events for each day for Health and Wellness app usage

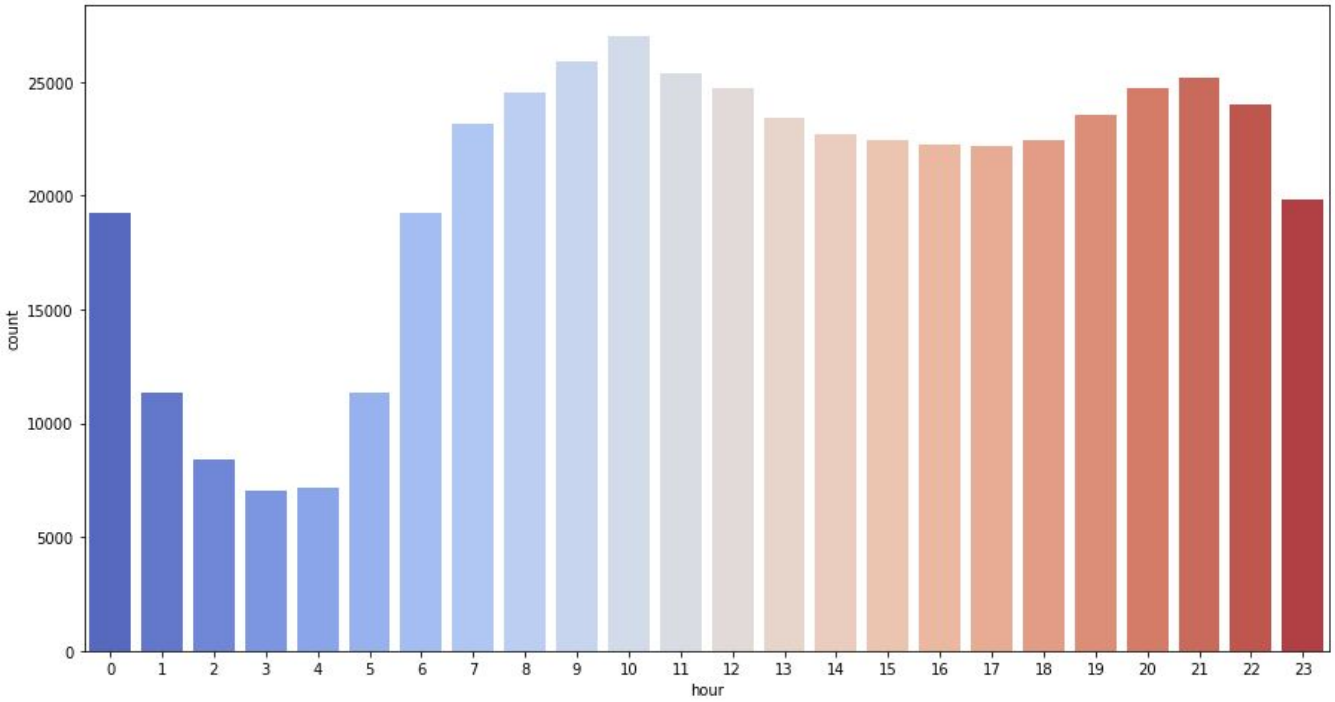


- Statistics for the above chart:

```
In [61]: # value counts per day
         device.week_day.value_counts()

Out[61]: Friday       74952
         Tuesday      73049
         Thursday     72004
         Wednesday    70026
         Saturday     66731
         Monday       66034
         Sunday       64451
         Name: week_day, dtype: int64

In [66]: # average usage per day
         device.week_day.value_counts().mean()

Out[66]: 69606.71428571429
```

We can see that Friday has the highest usage for Health and Wellness apps, Tuesday is the next, with Sunday being the least usage on Health and Wellness apps. The average daily usage is 69,606.7.

- Check hour usage: which hour has the most users? Count the number of events for each day for Health and Wellness app usage.



- Statistics for the above chart:

```
In [28]: # value counts per hour usage
         device.hour.value_counts()

Out[28]: 10    27033
         9     25926
         11    25384
         21    25179
         20    24736
         12    24709
         8     24526
         22    24016
         19    23563
         13    23393
         7     23145
         14    22710
         15    22435
         18    22411
         16    22258
         17    22169
         23    19844
         6     19260
         0     19223
         5     11366
         1     11346
         2      8423
         4      7162
         3      7030
         Name: hour, dtype: int64
```

```
In [31]: # average usage per hour
         device.hour.value_counts().mean()

Out[31]: 20301.958333333332
```

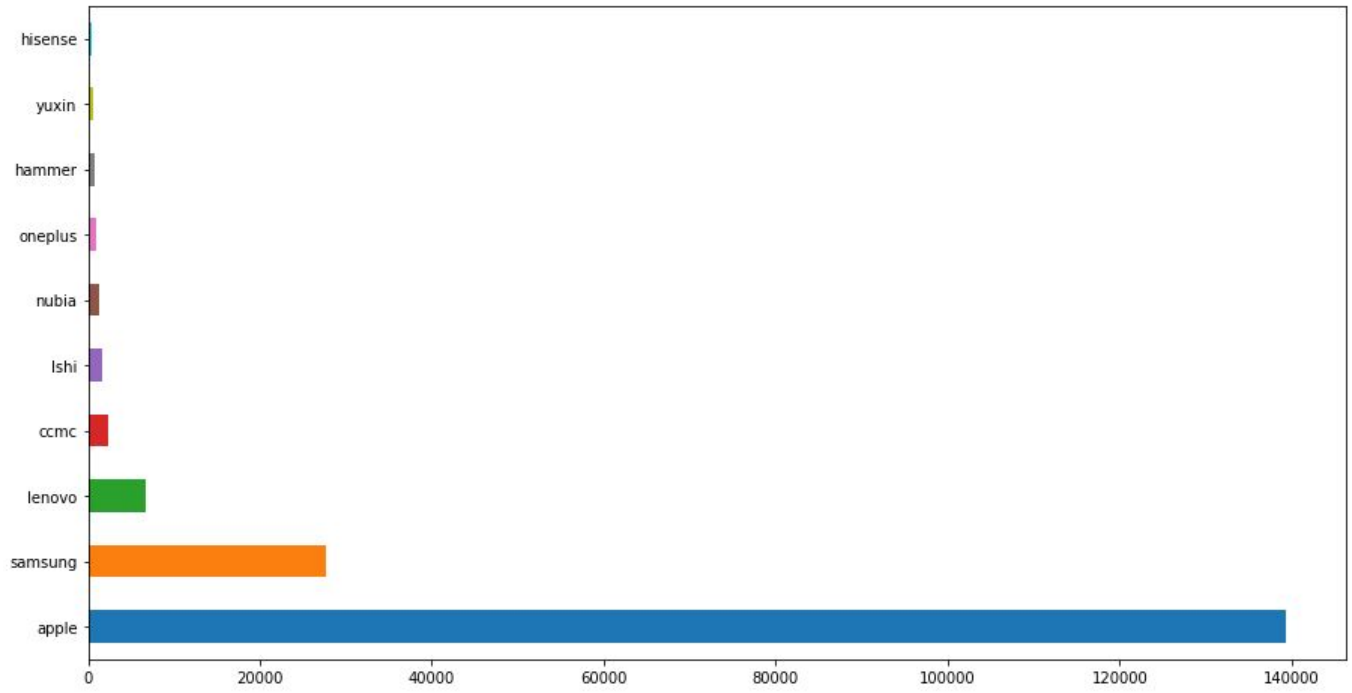
We can see that 9 to 10am, and 8pm to 9pm have the most usage for Health and Wellness apps. That's probably users are on the way to work in the morning and use the app often. Also, users tend

to do exercise at night after work. From midnight to dawn (0am to 6am), users use Health and Wellness apps the least, which is expected, since these are the time users are sleeping. The average count of usage per hour is about 20301.

Phone Brand: Which are the top phone brands for Health and Wellness apps users?

Logic of thinking: Merged app_event_category, event and gender_age_brand dataset together to get all the gender, age and phone brand information for Health and Wellness app category.

```
In [63]: # check the top 10 phone brand in using Health and Wellness apps
         user_demo.phone_brand.value_counts().head(10)

Out[63]: apple      139287
         samsung     27682
         lenovo       6788
         ccmc         2358
         lshi         1638
         nubia        1370
         oneplus       869
         hammer        706
         yuxin         573
         hisense       490
         Name: phone_brand, dtype: int64
```



Summary above shows that the top 10 phone brand that are using Health and Wellness Apps. No surprise, Apple and Samsung are the two most used phone brand. But Apple users are must more than Samsung.

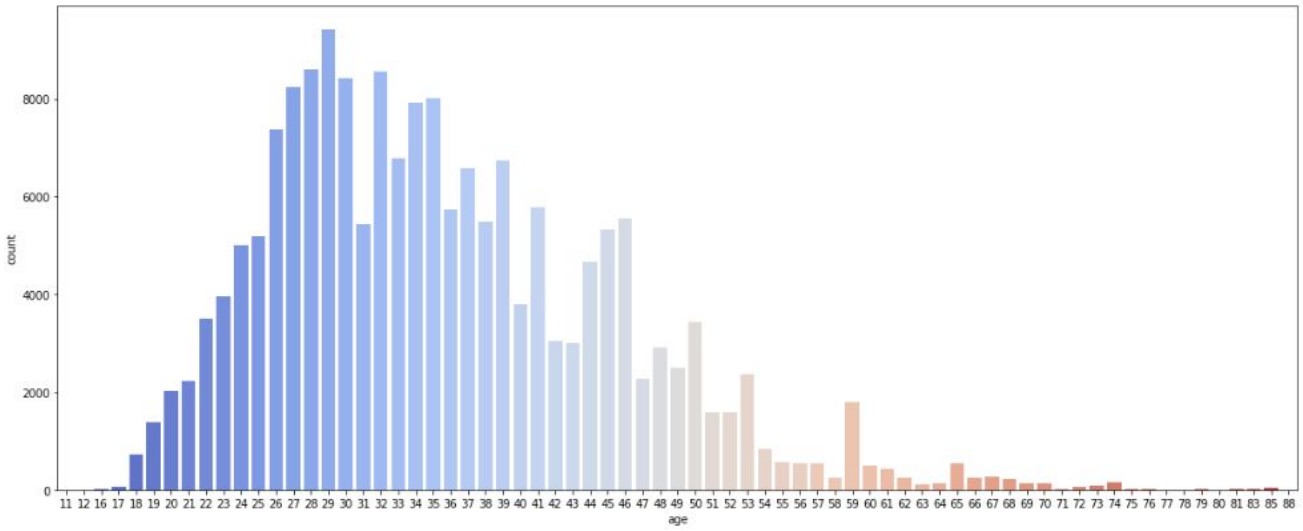
Age: which age use the Health and Wellness app more?

a. Average Age:

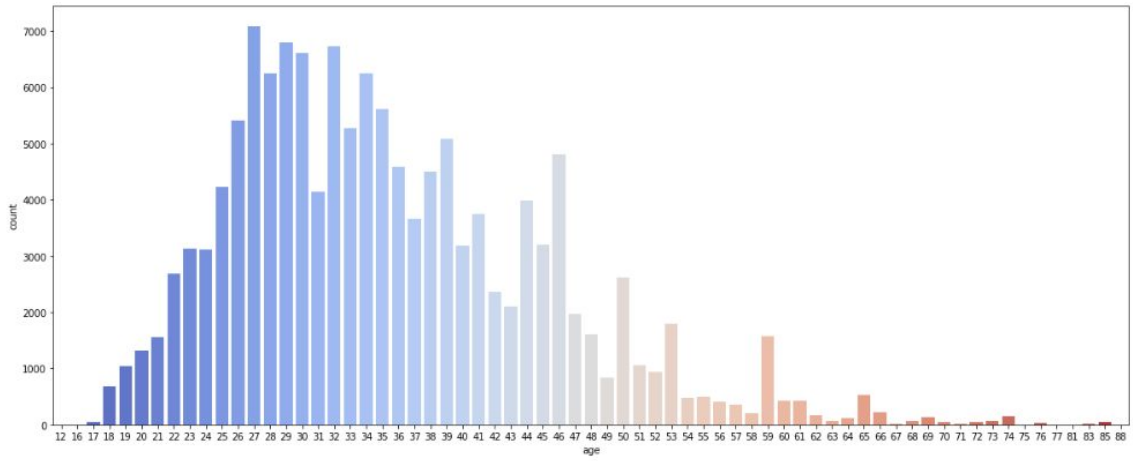
- Average age using the Health and Wellness app is: 35.6
- Average female age using the Health and Wellness app is: 36.1
- Average male age using the Health and Wellness app is: 35.4
- Not such difference in average age for using health and wellness app by gender. Mostly are age around 35.

```
In [96]: # average age using the Health and Wellness apps
         user_demo.age.mean()
Out[96]: 35.602278804282584

In [103]: user_demo[user_demo['gender'] == 'F'].age.mean()
Out[103]: 36.10626963239664

In [104]: user_demo[user_demo['gender'] == 'M'].age.mean()
Out[104]: 35.427852620135454
```

b. Age distribution in using the Health and Wellness app:
   - The graph below shows that the most users are around 26 to 36. There are more younger users than older users in the data.

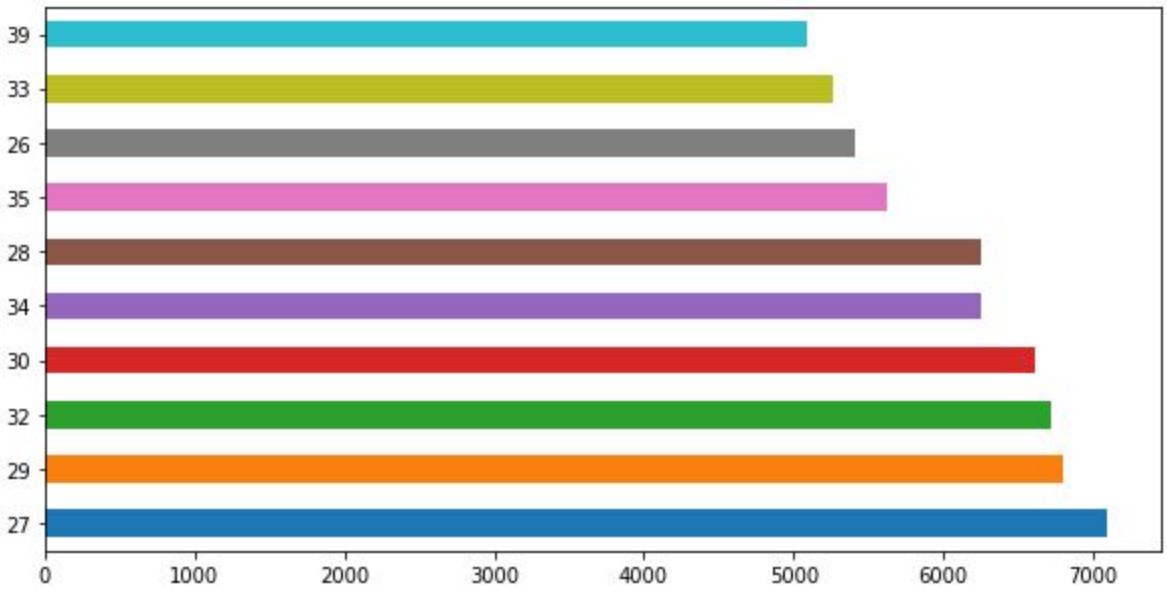


   - The graph below shows that the most male users are around 27 to 29 and 32 to 36. There are more younger male users than older male users in the data.

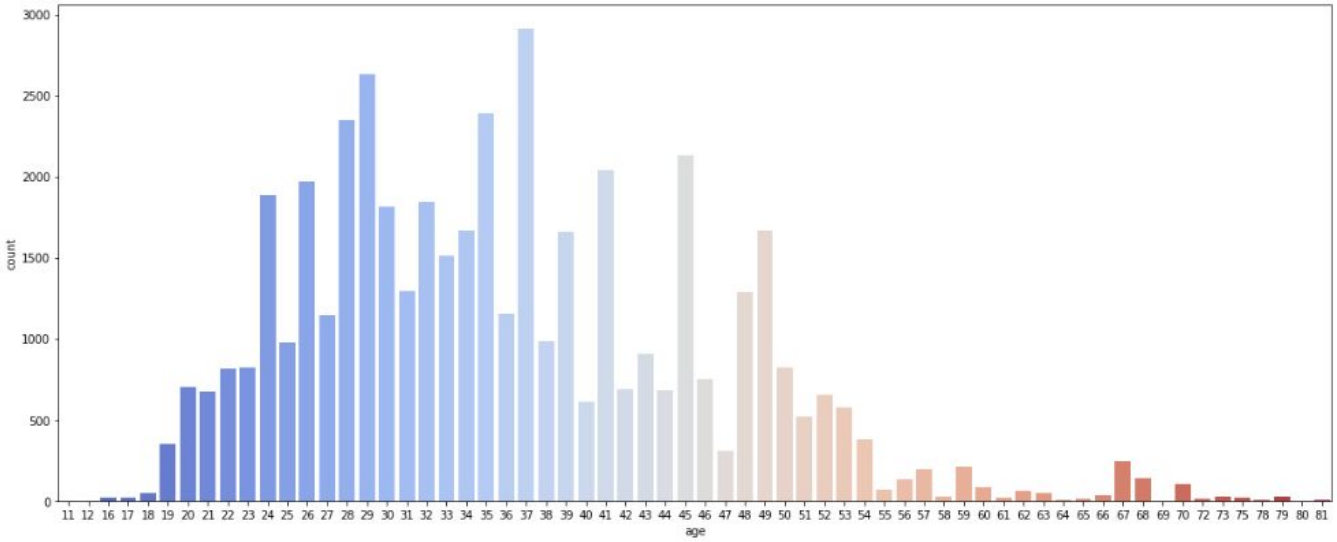


   - Top 10 age distribution for male users are:

```
In [145]: # top 10 male age group in using health and wellness app
          user_demo[user_demo['gender']=='M'].age.value_counts().head(10)

Out[145]: 27    7098
          29    6805
          32    6724
          30    6613
          34    6255
          28    6246
          35    5623
          26    5409
          33    5268
          39    5083
          Name: age, dtype: int64
```

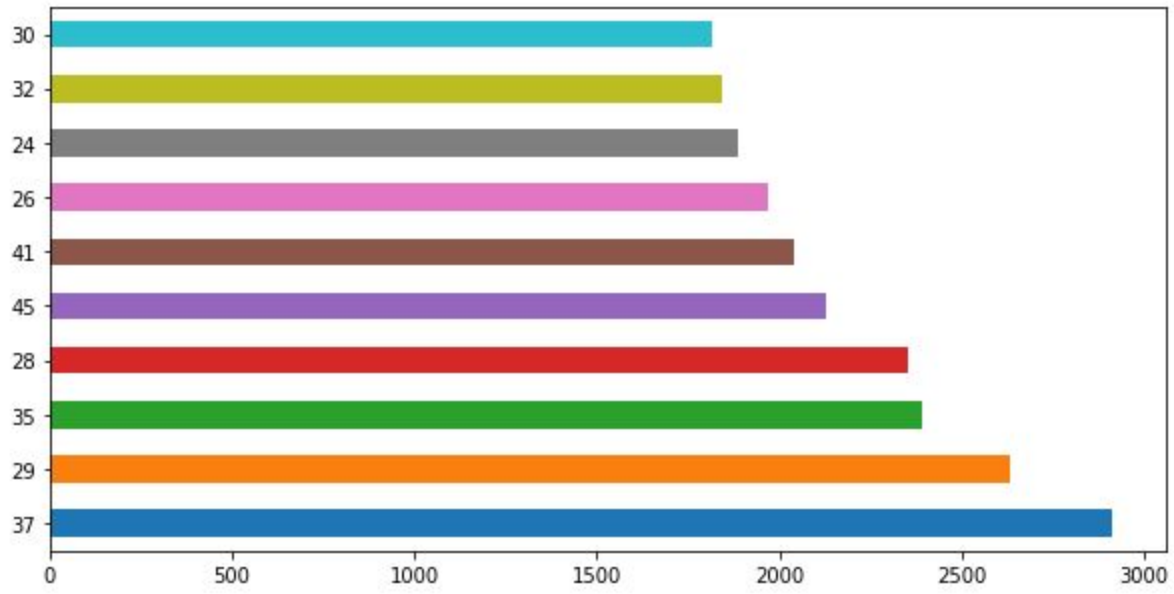


- The graph below also shows that there are more younger female than older female users in using Health and Wellness apps. But the distribution is more spread out than the male population.



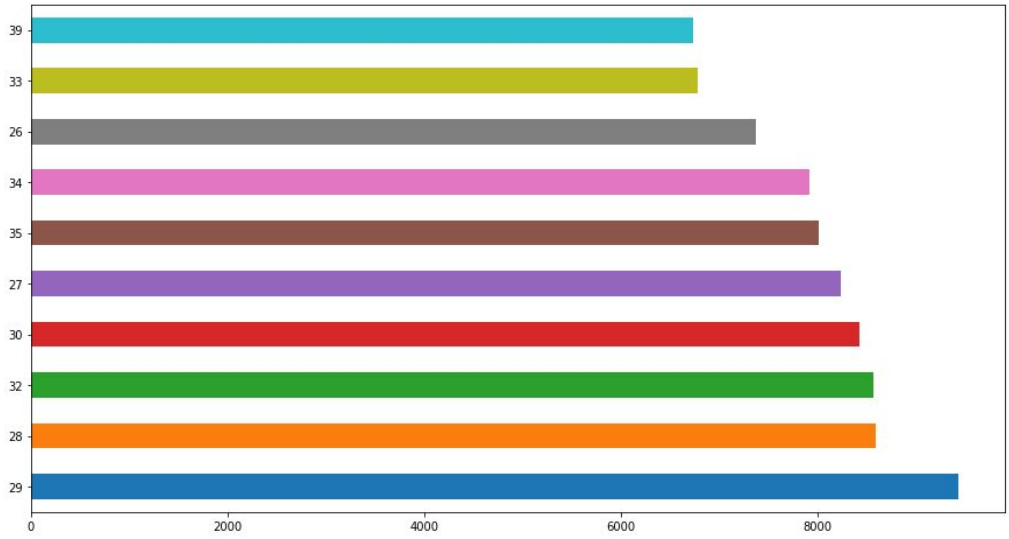
- Top 10 age distribution for female users are:

```
In [148]: # top 10 female age group in using health and wellness app
          user_demo[user_demo['gender']=='F'].age.value_counts().head(10)

Out[148]: 37    2915
          29    2632
          35    2393
          28    2351
          45    2128
          41    2040
          26    1968
          24    1885
          32    1844
          30    1815
          Name: age, dtype: int64
```



c. Top 10 age using the Health and Wellness app:

```
In [90]: # top 10 age using the Health and Wellness apps
         user_demo.age.value_counts().head(10)

Out[90]: 29    9437
         28    8597
         32    8568
         30    8428
         27    8241
         35    8016
         34    7922
         26    7377
         33    6780
         39    6741
         Name: age, dtype: int64
```

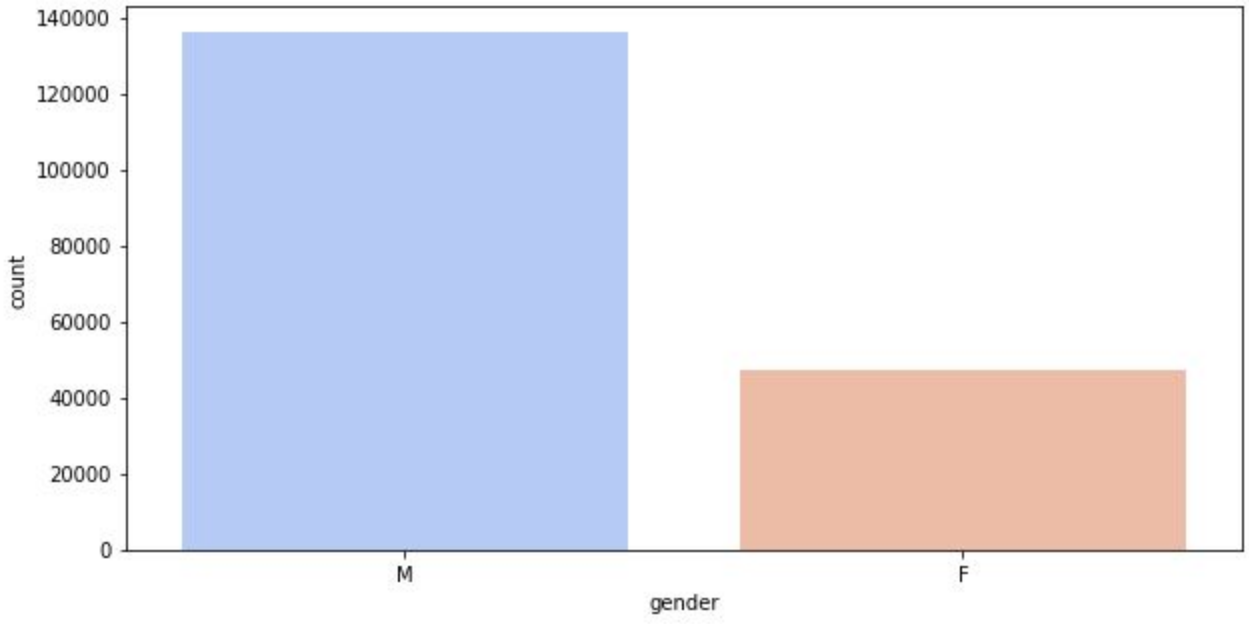


We can see that overall, younger users are more likely to use Health and Wellness apps. Users between 27 to 32 are the top 5 users who use Health and Wellness apps.

Gender: Which gender use the Health and Wellness more?

a. Gender distribution:

```
In [114]: # gender distribution in using the Health and Wellness apps
          user_demo.gender.value_counts()

Out[114]: M    136138
          F     47116
          Name: gender, dtype: int64
```

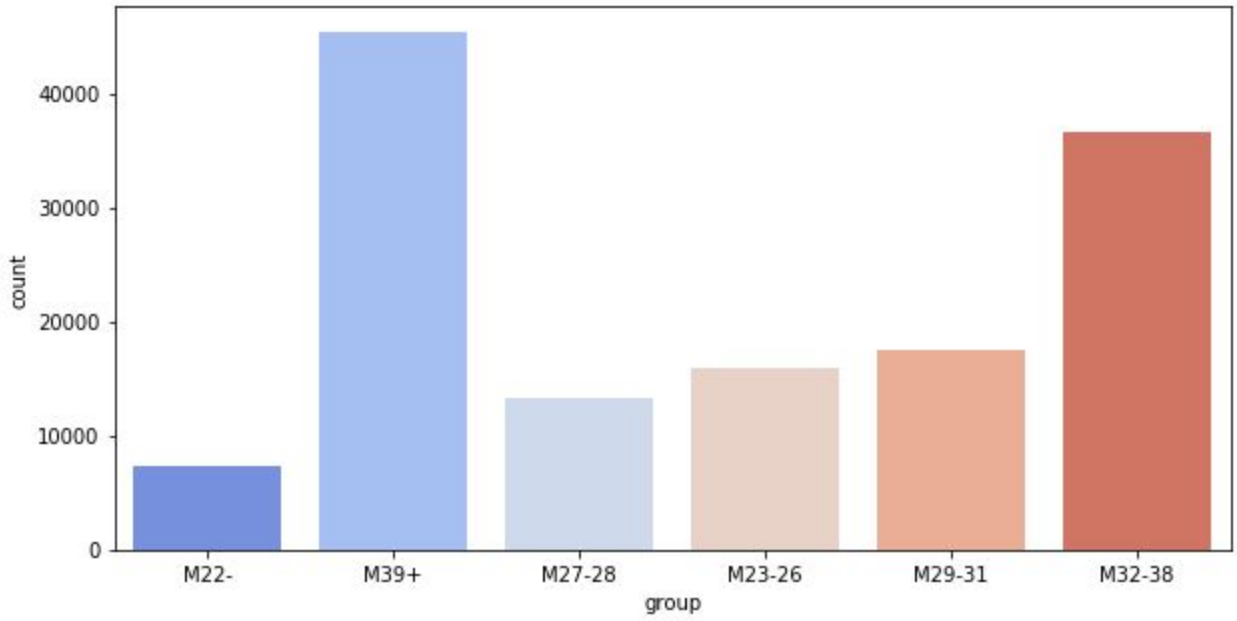


The dataset for male and female is very unbalanced. We have more male than female in the data for Health and Wellness apps. I'm not sure if the data is randomly sample or we just happened to sample more male than female. But assuming the data is randomly sampled, then we can see that male users are more active in using the Health and Wellness app.

Group: Which gender and age group combined use the Health and Wellness more? Assuming the data is randomly sampled.

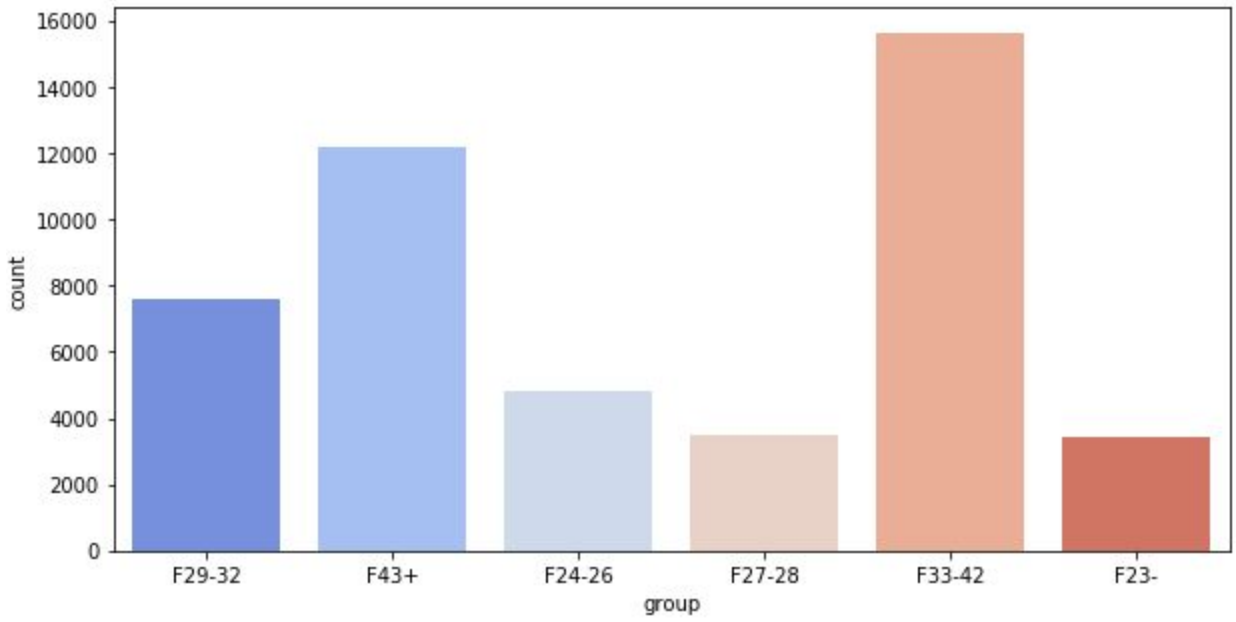
a. Check for Male and Age combined group: We can see that for male, age 39 or above are the users who use Health and Wellness apps more.

```
In [133]: # male and age combined distribution
          user_demo[user_demo['gender'] =='M'].group.value_counts()

Out[133]: M39+      45399
          M32-38    36607
          M29-31    17559
          M23-26    15873
          M27-28    13344
          M22-       7356
          Name: group, dtype: int64
```



b. Check for female and age combined group: We can see that for female, age 33 to 42 are the users who use Health and Wellness apps more.

```
In [135]: # female and age combined distribution
          user_demo[user_demo['gender'] =='F'].group.value_counts()

Out[135]: F33-42    15621
          F43+      12148
          F29-32     7582
          F24-26     4828
          F27-28     3494
          F23-       3443
          Name: group, dtype: int64
```



Gender and age combined market segments summary:

1. Overall, younger users are more likely to use the Health and Wellness apps.
2. Male are more active in using Health and Wellness apps than female.
3. If we want to target a large active group without separating gender, we can target age between 25 to 40. These covers most of the male and female gender population.
4. However, we also see that the age distribution between male and female are not the same. For male, the top age targettings could be from 29 to 40. For female, we can target age from 29 to 45.

2. Market Segments Should be Targeting Summary:

1. **Location:** China
2. **Day to target:** Weekday. Should focus on weekday. Weekend usage for Health and Wellness apps is less than weekday. Probably users are going out and having fun on weekends and not work on exercise.
3. **Hours to target:** Morning from 9 - 10, and evening from 8 - 9. These are time users user the app while they are going to work, and also users like to do exercise after work in evening.
4. **Phone Brand:** Top phone brand in using health and wellness apps are Apple and Samsung, with Apple much more than Samsung. Should target on IOS and Android system, but should prioritize the IOS development since these are the major users.
5. **Age:** Younger users, with average age = 35.6. The top age range in using Health and Wellness app, regardless gender is between 25 and 40. These are the primary users in health and wellness app.
6. **Gender:** Male. Male users tend to use health and wellness app more than female users.

3. The client is interested in partnering with apps outside the Health and Wellness space to drive cross promotion. Where are the biggest opportunities? Please create some short bullet points to support the recommendation in both data science and business contexts.

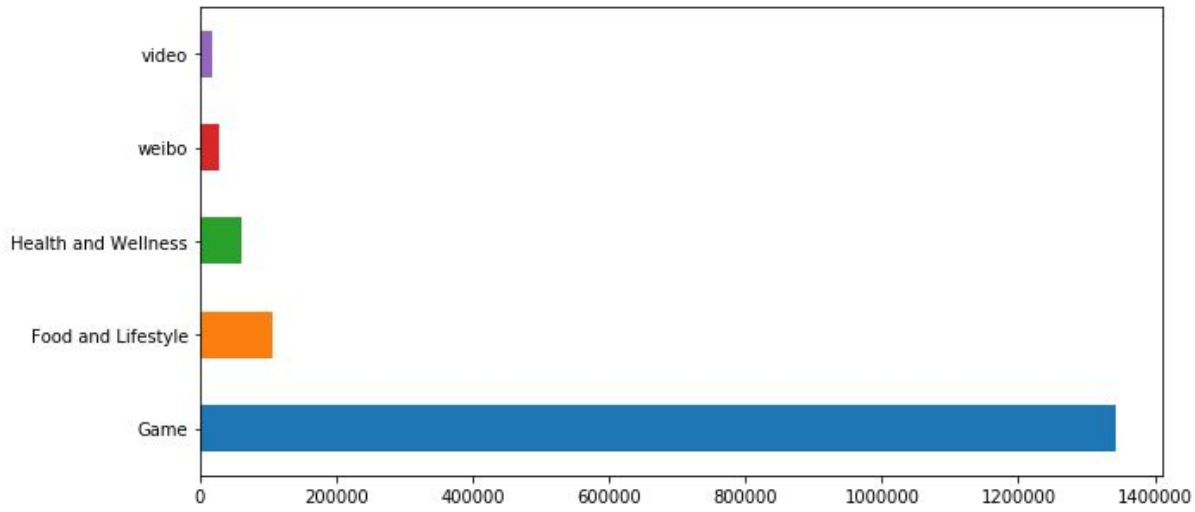
Logic of thinking: Since we've figured out which market segments are the health and wellness app users belong to in step 2. We can use those market segments to find what other app categories those market segments population would like to use. Then we can partner with those category to drive cross promotion.

That means we are going to find apps outside the Health and Wellness space that have the most users in the following criteria:

- For weekdays:
    - Location in China
    - Iphone, samsung users
    - Usually use apps between 9 - 12 morning time, and 8 - 9 evening time
    - Younger users, usually between 25 to 40
- For weekends:
    - Location in China
    - Iphone, samsung users
    - Younger age,usually between 25 to 40
    - Male prefer

1. Find top app categories outside health and wellness app in China:

```
In [164]: china.app_category.value_counts().head(5)

Out[164]: Game                   1343905
          Food and Lifestyle      105335
          Health and Wellness      61098
          weibo                    26748
          video                    18090
          Name: app_category, dtype: int64
```



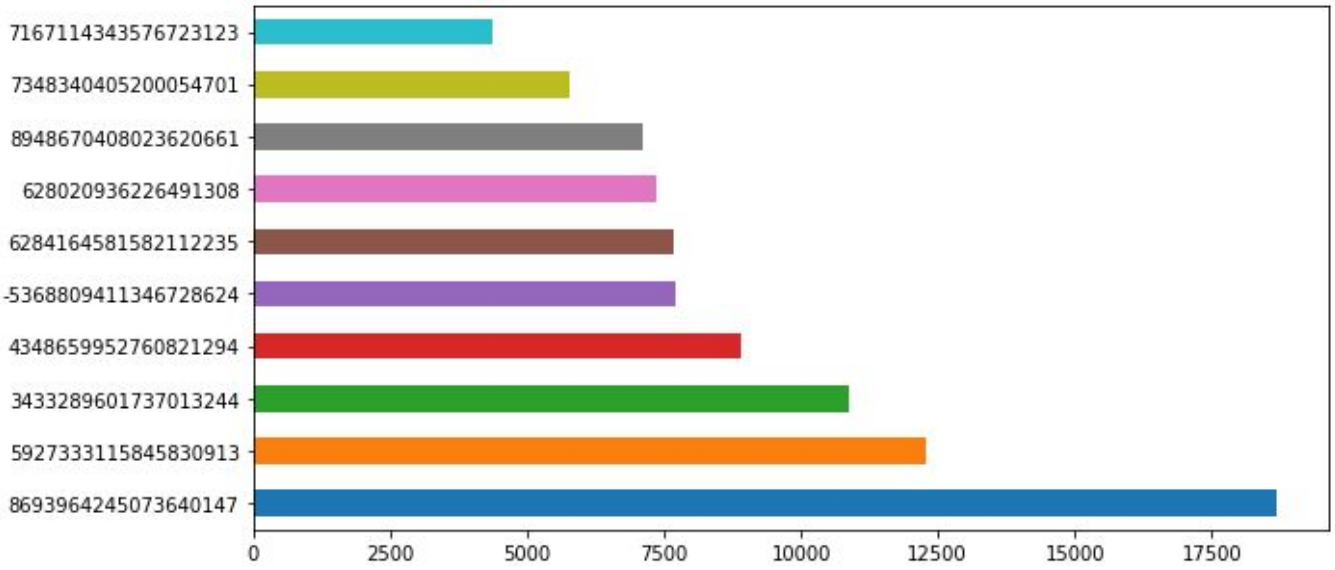
The above 5 categories are the top apps being used in China. Beside Health and Wellness, app category we can consider are Game, Food and Lifestyle, weibo and video. But we see Game has much more users than any other category. Let's focus on **game** category to find apps that have the most opportunities.

2. Find apps in game category that have the biggest opportunities to partner with on weekdays to drive cross-promotion. I'm looking for apps:

- In Game category
- Being played during weekdays

- Mostly being played between 9 - 12 am, and 8 - 9pm
- Phone brands are in apple and samsung
- Age range from 25 to 40
- Below are the top 10 apps in these criteria and have the most opportunities to partner with the health and wellness apps

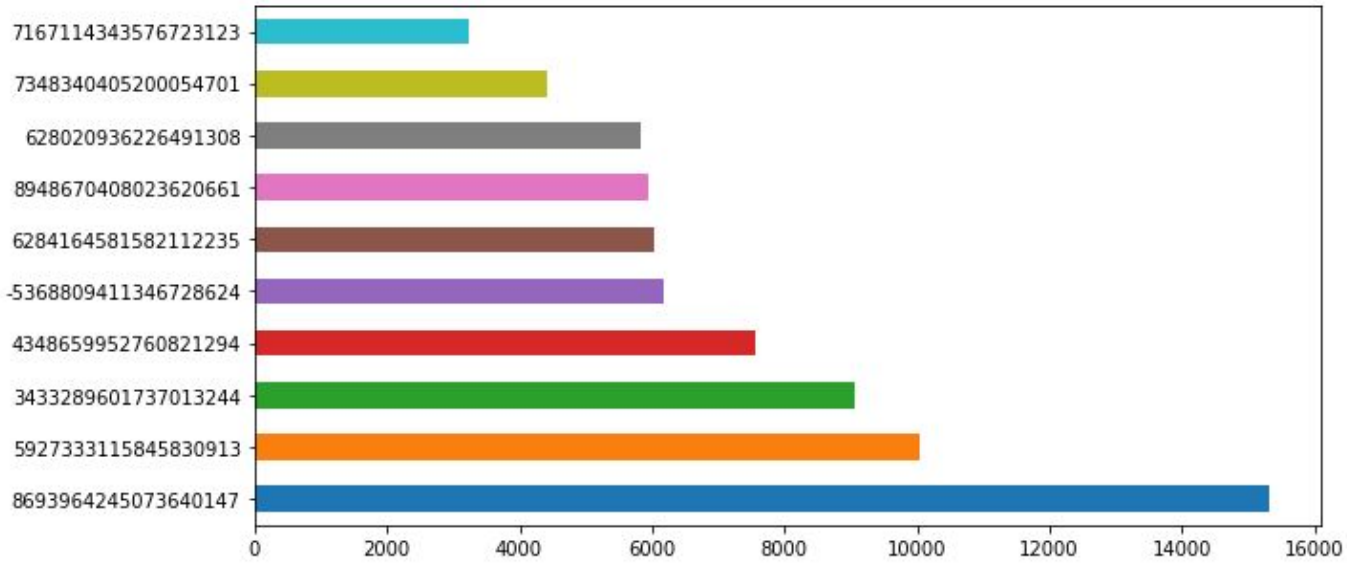
```
In [220]: df3.app_id.value_counts().head(10)

Out[220]: 8693964245073640147     18695
          5927333115845830913     12286
          3433289601737013244     10890
          4348659952760821294      8914
          -5368809411346728624     7710
          6284164581582112235      7672
          628020936226491308       7345
          8948670408023620661      7111
          7348340405200054701      5778
          7167114343576723123      4353
          Name: app_id, dtype: int64
```



2. Find apps in game category that have the biggest opportunities to partner with on weekends to drive cross-promotion. I'm looking for apps:

- In Game category
- Being played during weekends
- Phone brands are in apple and samsung
- Age range from 25 to 40
- Male prefer
- Below are the top 10 apps in these criteria and have the most opportunities to partner with the health and wellness apps

```
Out[236]: 8693964245073640147     15323
          5927333115845830913     10044
          3433289601737013244      9052
          4348659952760821294      7548
          -5368809411346728624     6189
          6284164581582112235      6045
          8948670408023620661      5943
          628020936226491308       5827
          7348340405200054701      4410
          7167114343576723123      3231
          Name: app_id, dtype: int64
```



Compared both weekdays and weekends data, I found out the top 10 most played game. that have the similar market segments compared to the health and wellness are the same. They are the following app_ids:

```
8693964245073640147
5927333115845830913
3433289601737013244
4348659952760821294
-5368809411346728624
6284164581582112235
8948670408023620661
628020936226491308
7348340405200054701
7167114343576723123
```

These are the Game apps that have the biggest opportunities to drive cross-promotion with the Health and Wellness apps because they share the similar market segments and these apps currently have the most market share.

I also checked the Food and Lifestyle category to see if there are any apps that have bigger market shares than the ones I found in the game categories that can be used to help drive cross-promotion. But even

weekdays data in “Food and Lifestyle” category won’t exceed the game market share. See below. I think the biggest opportunity is to focus on Game category, and the 10 app_ids identified above.

Below are the top 10 Food and Lifestyle apps that have similar market segments as the health and wellness space.

```
Out[245]: 8693964245073640147     15323
          5927333115845830913     10044
          3433289601737013244      9052
          4348659952760821294      7548
          -5368809411346728624     6189
          6284164581582112235      6045
          8948670408023620661      5943
          628020936226491308       5827
          7348340405200054701      4410
          7167114343576723123      3231
          Name: app_id, dtype: int64
```