Apple App Store Games analysis and visualization project

For this project, I chose a dataset from kaggle.com that collects data about various app store games and the records of some of their characteristics. While exploring and cleaning the data, I have posed some questions that the data set might have an answer to:

- What are the frequencies of the average user ratings and the prices?
- What is the count of different age ratings?
- What are the average game sizes of the two most common primary genres?
- What is the difference between the two average sizes?

In [1]:
```
import altair as alt
import pandas as pd
```

In [2]:
```
df = pd.read_csv('appstore_games.csv')
```

In [3]:
```
df.columns
```

Out[3]:
```
Index(['URL', 'ID', 'Name', 'Subtitle', 'Icon URL', 'Average User Rating',
       'User Rating Count', 'Price', 'In-app Purchases', 'Description',
       'Developer', 'Age Rating', 'Languages', 'Size', 'Primary Genre',
       'Genres', 'Original Release Date', 'Current Version Release Date'],
      dtype='object')
```

In [4]:
```
game_df = df.fillna(method = 'ffill')
```

In [5]:
```
df2 = game_df['Average User Rating']
df2
```

Out[5]:
```
0        4.0
1        3.5
2        3.0
3        3.5
4        3.5
        ...
17002    5.0
17003    5.0
17004    5.0
17005    5.0
17006    5.0
Name: Average User Rating, Length: 17007, dtype: float64
```

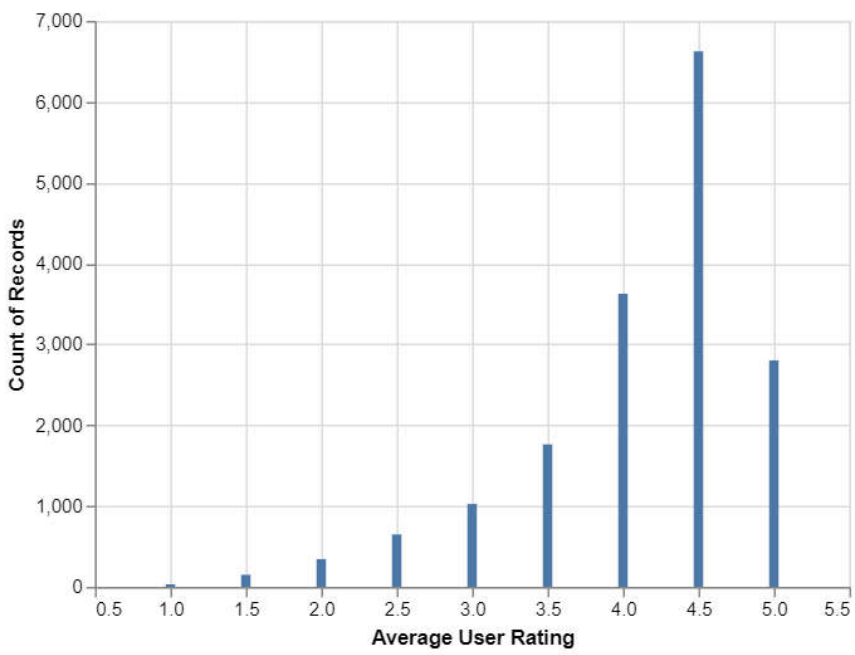
1. What are the frequencies of the average user ratings and the prices?

In [6]:
```
alt.data_transformers.disable_max_rows()
```

Out[6]:
```
DataTransformerRegistry.enable('default')
```

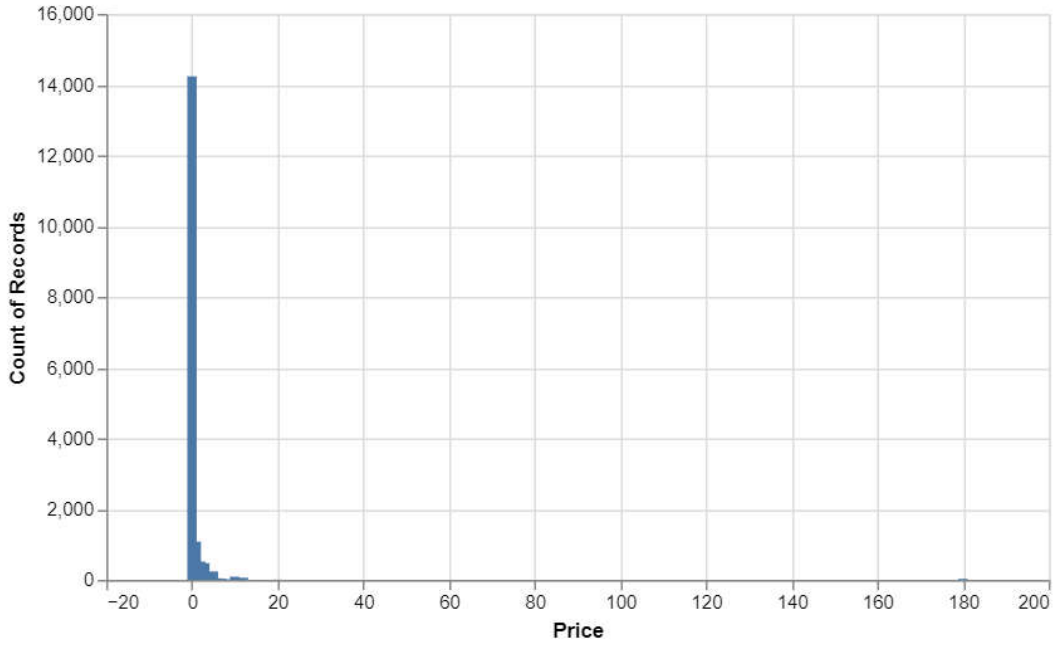
In [7]:
```
average_user_rating_count = alt.Chart(game_df).mark_bar().encode(x= 'Average User Rating', y= 'count()', tooltip='count()')
average_user_rating_count
```

Out[7]:



In [8]:
```
price_count = alt.Chart(game_df, width = 500).mark_bar().encode(x= 'Price', y= 'count()', tooltip = 'count()')
price_count
```

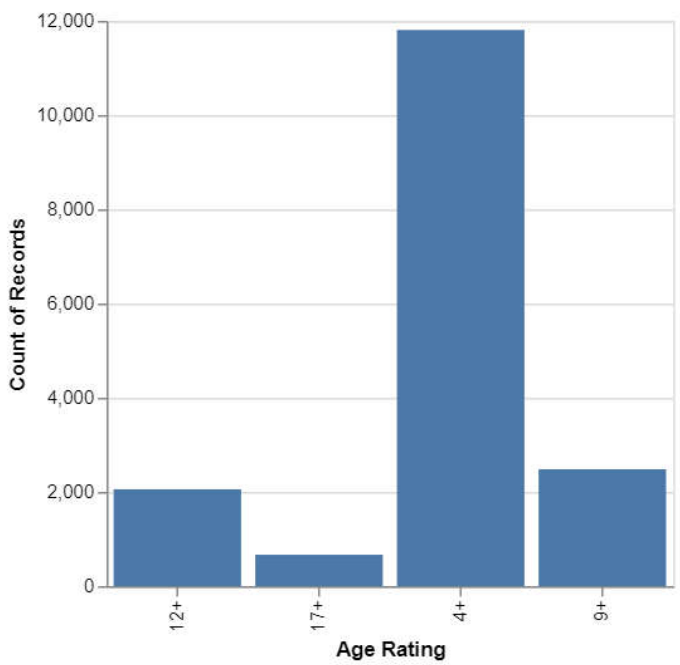
Out[8]:



1. What is the count of different age ratings?

In [9]:
```
age_rating_count = alt.Chart(game_df, width = 300).mark_bar().encode(x = 'Age Rating', y= 'count()', tooltip = 'count()')
age_rating_count
```

Out[9]:



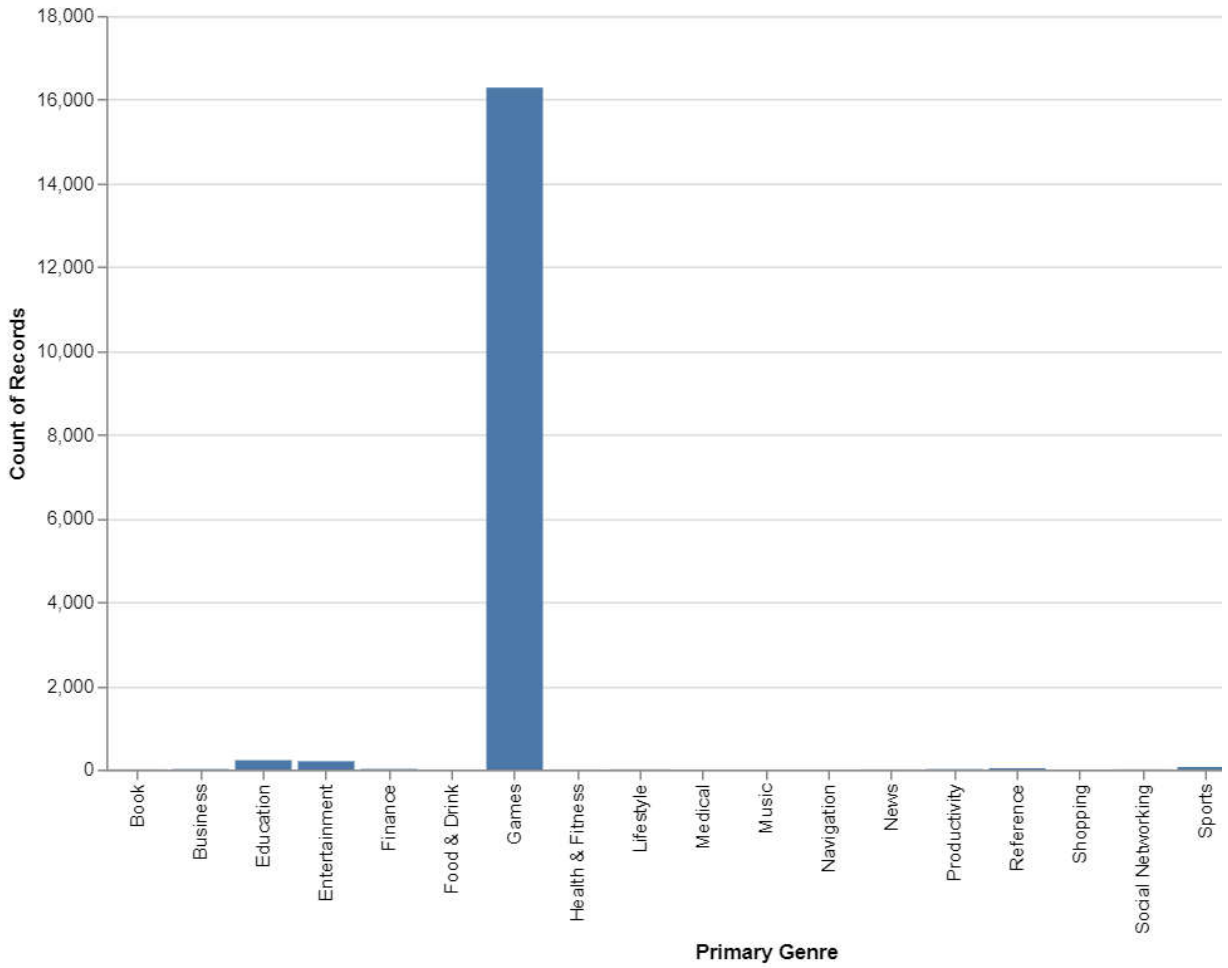
1. What are the average game sizes of the two most common primary genres?

In [10]:
```
game_df['Primary Genre'].unique()
```

Out[10]:
```
array(['Games', 'Entertainment', 'Finance', 'Sports', 'Reference',
       'Medical', 'Education', 'Utilities', 'Book', 'Travel',
       'Productivity', 'Lifestyle', 'Business', 'News',
       'Social Networking', 'Health & Fitness', 'Music', 'Stickers',
       'Food & Drink', 'Shopping', 'Navigation'], dtype=object)
```

In [11]:
```
primary_genre_count = alt.Chart(game_df, width=700, height = 400).mark_bar().encode(
x='Primary Genre', y= 'count()', tooltip = 'count()')
primary_genre_count
```

Out[11]:



In [12]:
```
primary_genre_table = game_df.groupby('Primary Genre').agg(mean_size = ('Size', 'mean')).reset_index()
primary_genre_table
```

Out[12]:

	Primary Genre	mean_size
0	Book	6.860493e+07
1	Business	7.045702e+07
2	Education	1.049152e+08
3	Entertainment	7.255211e+07
4	Finance	6.159167e+07
5	Food & Drink	4.599347e+07
6	Games	1.174199e+08
7	Health & Fitness	2.892120e+08
8	Lifestyle	6.701885e+07
9	Medical	1.078185e+07
10	Music	4.884446e+07
11	Navigation	1.675878e+07
12	News	4.462870e+07
13	Productivity	4.995284e+07
14	Reference	4.825379e+07
15	Shopping	5.584589e+07
16	Social Networking	8.421103e+07
17	Sports	4.623130e+07
18	Stickers	3.908890e+06
19	Travel	1.210470e+07
20	Utilities	8.299363e+07

1. What is the difference between the average sizes of the "Games" and the "Education" genre?

In [13]:
```
primary_genre_series = pd.Series(primary_genre_table['mean_size'].values, index = pr
imary_genre_table['Primary Genre'])
primary_genre_series
```

Out[13]:
```
Primary Genre
Book                 6.860493e+07
Business             7.045702e+07
Education            1.049152e+08
Entertainment        7.255211e+07
Finance              6.159167e+07
Food & Drink         4.599347e+07
Games                1.174199e+08
Health & Fitness     2.892120e+08
Lifestyle            6.701885e+07
Medical              1.078185e+07
Music                4.884446e+07
Navigation           1.675878e+07
News                 4.462870e+07
Productivity         4.995284e+07
Reference            4.825379e+07
Shopping             5.584589e+07
Social Networking    8.421103e+07
Sports               4.623130e+07
Stickers             3.908890e+06
Travel               1.210470e+07
Utilities            8.299363e+07
dtype: float64
```

In [14]:
```
size_difference = primary_genre_series['Games'] - primary_genre_series['Education']
size_difference
```

Out[14]:
```
12504752.524923295
```