In [4]:
```
library(tidyverse)
library(dplyr)
library(readxl)
data <- read_excel("D:/Data.xlsx")
head(data)
summary(data)
glimpse(data)
```

Movie	Year	Ratings	Genre	Gross	Budget	Screens	Sequel	Sentiment	Views
13 Sins	2014	6.3	8	9.13e+03	4.0e+06	45	1	0	3280543
22 Jump Street	2014	7.1	1	1.92e+08	5.0e+07	3306	2	2	583289
3 Days to Kill	2014	6.2	1	3.07e+07	2.8e+07	2872	1	0	304861
300: Rise of an Empire	2014	6.3	1	1.06e+08	1.1e+08	3470	2	0	452917
A Haunted House 2	2014	4.7	8	1.73e+07	3.5e+06	2310	2	0	3145573
A Long Way Off	2014	4.6	3	2.90e+04	5.0e+05	NA	1	0	91137

```
   Movie               Year         Ratings          Genre       
 Length:231         Min.   :2014   Min.   :3.100   Min.   : 1.000  
 Class :character   1st Qu.:2014   1st Qu.:5.800   1st Qu.: 1.000  
 Mode  :character   Median :2014   Median :6.500   Median : 3.000  
                    Mean   :2014   Mean   :6.442   Mean   : 5.359  
                    3rd Qu.:2015   3rd Qu.:7.100   3rd Qu.: 8.000  
                    Max.   :2015   Max.   :8.700   Max.   :15.000  

     Gross              Budget             Screens         Sequel      
 Min.   :     2470   Min.   :    70000   Min.   :   2   Min.   :1.000  
 1st Qu.: 10300000   1st Qu.:  9000000   1st Qu.: 449   1st Qu.:1.000  
 Median : 37400000   Median : 28000000   Median :2777   Median :1.000  
 Mean   : 68066033   Mean   : 47921730   Mean   :2209   Mean   :1.359  
 3rd Qu.: 89350000   3rd Qu.: 65000000   3rd Qu.:3372   3rd Qu.:1.000  
 Max.   :643000000   Max.   :250000000   Max.   :4324   Max.   :7.000  
                     NA's   :1           NA's   :10                    
   Sentiment          Views               Likes            Dislikes      
 Min.   :-38.00   Min.   :     698   Min.   :     1   Min.   :    0.0  
 1st Qu.:  0.00   1st Qu.:  623302   1st Qu.:  1776   1st Qu.:  105.5  
 Median :  0.00   Median : 2409338   Median :  6096   Median :  341.0  
 Mean   :  2.81   Mean   : 3712851   Mean   : 12732   Mean   :  679.1  
 3rd Qu.:  5.50   3rd Qu.: 5217380   3rd Qu.: 15248   3rd Qu.:  697.5  
 Max.   : 29.00   Max.   :32626778   Max.   :370552   Max.   :13960.0  

    Comments       Aggregate Followers
 Min.   :    0.0   Min.   :    1066  
 1st Qu.:  248.5   1st Qu.:  183025  
 Median :  837.0   Median : 1052600  
 Mean   : 1825.7   Mean   : 3038193  
 3rd Qu.: 2137.0   3rd Qu.: 3694500  
 Max.   :38363.0   Max.   :31030000  
                   NA's   :35        
Observations: 231
Variables: 14
$ Movie                <chr> "13 Sins", "22 Jump Street", "3 Days to Kill"...
$ Year                 <dbl> 2014, 2014, 2014, 2014, 2014, 2014, 2014, 201...
$ Ratings              <dbl> 6.3, 7.1, 6.2, 6.3, 4.7, 4.6, 6.1, 7.1, 6.5, ...
$ Genre                <dbl> 8, 1, 1, 1, 8, 3, 8, 1, 10, 8, 1, 8, 15, 8, 3...
$ Gross                <dbl> 9.13e+03, 1.92e+08, 3.07e+07, 1.06e+08, 1.73e...
$ Budget               <dbl> 4.00e+06, 5.00e+07, 2.80e+07, 1.10e+08, 3.50e...
$ Screens              <dbl> 45, 3306, 2872, 3470, 2310, NA, 3158, 818, 27...
$ Sequel               <dbl> 1, 2, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, ...
$ Sentiment            <dbl> 0, 2, 0, 0, 0, 0, 0, 2, 3, 0, 4, 0, 0, 0, 0, ...
$ Views                <dbl> 3280543, 583289, 304861, 452917, 3145573, 911...
$ Likes                <dbl> 4632, 3465, 328, 2429, 12163, 112, 9595, 2207...
$ Dislikes             <dbl> 425, 61, 34, 132, 610, 7, 419, 197, 419, 532,...
$ Comments             <dbl> 636, 186, 47, 590, 1082, 1, 1020, 593, 382, 7...
$ `Aggregate Followers` <dbl> 1120000, 12350000, 483000, 568000, 1923800, 3...
```

In [8]:
```r
high_rated_movies <- data %>%
  filter(Ratings > 8)
high_rated_movies
```

Movie	Year	Ratings	Genre	Gross	Budget	Screens	Sequel	Sentiment	Viev
Gone Girl	2014	8.2	3	1.68e+08	6.10e+07	3014	1	-11	3960
Guardians of the Galaxy	2014	8.1	1	3.33e+08	1.70e+08	4080	1	0	13135
Interstellar	2014	8.7	2	1.88e+08	1.65e+08	3561	1	2	54217
The Imitation Game	2014	8.1	9	9.11e+07	1.40e+07	747	1	9	30478
Whiplash	2014	8.6	3	1.31e+07	3.30e+06	42	1	2	77502
Wild Tales	2014	8.2	8	3.08e+06	3.30e+06	4	1	7	6971
X-Men: Days of Future Past	2014	8.1	1	2.34e+08	2.00e+08	3996	7	3	228
Inside Out	2015	8.6	12	3.45e+08	1.75e+08	3946	1	2	14389
Mad Max: Fury Road	2015	8.3	1	1.53e+08	1.50e+08	3702	4	-4	27323
Straight Outta Compton	2015	8.3	9	1.35e+08	2.80e+07	2757	1	5	8489
Me and Earl and the Dying Girl	2015	8.2	8	6.74e+06	8.00e+06	34	1	15	40322

In [9]:
```
distinct_genres <- data %>%
  distinct(Genre)
distinct_genres
```

Genre
8
1
3
10
15
12
9
2
7
6
4

In [10]:
```
data <- data %>%
  mutate(Gross_in_Millions = Gross / 1e6)
head(data)
```

Movie	Year	Ratings	Genre	Gross	Budget	Screens	Sequel	Sentiment	Views
13 Sins	2014	6.3	8	9.13e+03	4.0e+06	45	1	0	3280543
22 Jump Street	2014	7.1	1	1.92e+08	5.0e+07	3306	2	2	583289
3 Days to Kill	2014	6.2	1	3.07e+07	2.8e+07	2872	1	0	304861
300: Rise of an Empire	2014	6.3	1	1.06e+08	1.1e+08	3470	2	0	452917
A Haunted House 2	2014	4.7	8	1.73e+07	3.5e+06	2310	2	0	3145573
A Long Way Off	2014	4.6	3	2.90e+04	5.0e+05	NA	1	0	91137

In [13]:
```
movie_financials <- data %>%
  transmute(Movie, Gross_in_Millions, Budget)
head(movie_financials)
```

Movie	Gross_in_Millions	Budget
13 Sins	9.13e-03	4.0e+06
22 Jump Street	1.92e+02	5.0e+07
3 Days to Kill	3.07e+01	2.8e+07
300: Rise of an Empire	1.06e+02	1.1e+08
A Haunted House 2	1.73e+01	3.5e+06
A Long Way Off	2.90e-02	5.0e+05

In [15]:
```
genre_summary <- data %>%
  group_by(Genre) %>%
  summarise(
    avg_gross = mean(Gross, na.rm = TRUE),
    avg_budget = mean(Budget, na.rm = TRUE)
  )
genre_summary
```

Genre	avg_gross	avg_budget
1	114495349	83558462
2	124040833	85875000
3	31475425	22875715
4	1210000	50000000
6	25326667	31000000
7	3355	3650000
8	43530902	25518302
9	45156692	21384615
10	23694500	21879167
12	134761538	104000000
15	29356000	11259500

In [16]:
```
data <- data %>%
  mutate(Gross_in_Millions = Gross / 1e6)


head(data)
```

Movie	Year	Ratings	Genre	Gross	Budget	Screens	Sequel	Sentiment	Views
13 Sins	2014	6.3	8	9.13e+03	4.0e+06	45	1	0	3280543
22 Jump Street	2014	7.1	1	1.92e+08	5.0e+07	3306	2	2	583289
3 Days to Kill	2014	6.2	1	3.07e+07	2.8e+07	2872	1	0	304861
300: Rise of an Empire	2014	6.3	1	1.06e+08	1.1e+08	3470	2	0	452917
A Haunted House 2	2014	4.7	8	1.73e+07	3.5e+06	2310	2	0	3145573
A Long Way Off	2014	4.6	3	2.90e+04	5.0e+05	NA	1	0	91137

In [17]:
```
movie_financials <- data %>%
  transmute(Movie, Gross_in_Millions, Budget)


head(movie_financials)
```

Movie	Gross_in_Millions	Budget
13 Sins	9.13e-03	4.0e+06
22 Jump Street	1.92e+02	5.0e+07
3 Days to Kill	3.07e+01	2.8e+07
300: Rise of an Empire	1.06e+02	1.1e+08
A Haunted House 2	1.73e+01	3.5e+06
A Long Way Off	2.90e-02	5.0e+05

In [20]:
```
distinct_genres <- data %>%
  distinct(Genre)
distinct_genres
```

Genre
8
1
3
10
15
12
9
2
7
6
4

In [21]:
```
data <- data %>%
  mutate(
    Gross_in_Millions = Gross / 1e6,
    Gross_Per_Screen = Gross / Screens,
    Profit = Gross - Budget
  )


head(data)
```

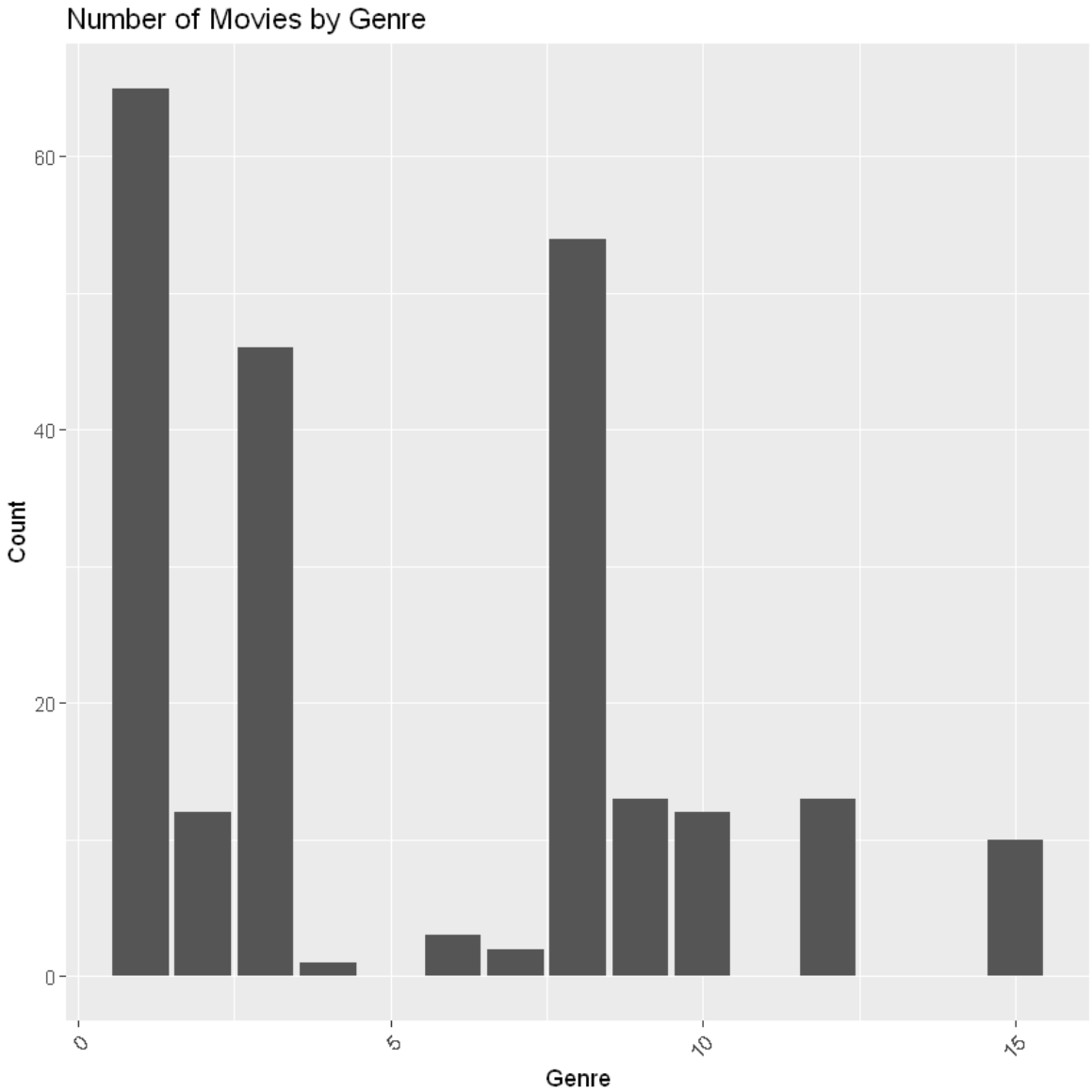
Movie	Year	Ratings	Genre	Gross	Budget	Screens	Sequel	Sentiment	Views
13 Sins	2014	6.3	8	9.13e+03	4.0e+06	45	1	0	3280543
22 Jump Street	2014	7.1	1	1.92e+08	5.0e+07	3306	2	2	583289
3 Days to Kill	2014	6.2	1	3.07e+07	2.8e+07	2872	1	0	304861
300: Rise of an Empire	2014	6.3	1	1.06e+08	1.1e+08	3470	2	0	452917
A Haunted House 2	2014	4.7	8	1.73e+07	3.5e+06	2310	2	0	3145573
A Long Way Off	2014	4.6	3	2.90e+04	5.0e+05	NA	1	0	91137

In [22]:
```
movie_financials <- data %>%
  transmute(Movie, Gross_in_Millions, Budget, Profit, Gross_Per_Screen)
```

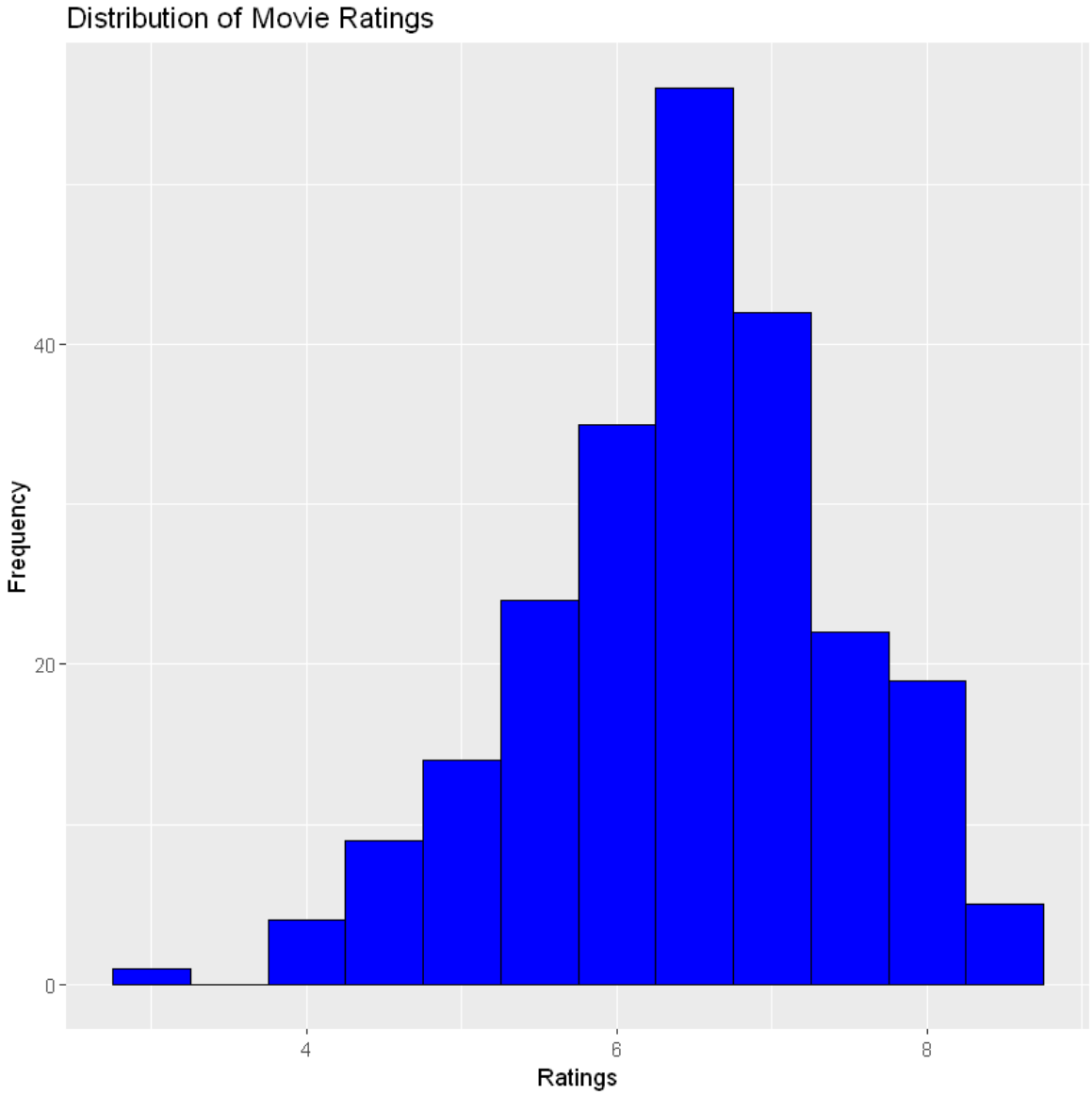
```
head(movie_financials)
```

Movie	Gross_in_Millions	Budget	Profit	Gross_Per_Screen
13 Sins	9.13e-03	4.0e+06	-3990870	202.8889
22 Jump Street	1.92e+02	5.0e+07	142000000	58076.2250
3 Days to Kill	3.07e+01	2.8e+07	2700000	10689.4150
300: Rise of an Empire	1.06e+02	1.1e+08	-4000000	30547.5504
A Haunted House 2	1.73e+01	3.5e+06	13800000	7489.1775
A Long Way Off	2.90e-02	5.0e+05	-471000	NA

In [23]:
```
ggplot(data, aes(x = Genre)) +
  geom_bar() +
  labs(title = "Number of Movies by Genre",
       x = "Genre",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

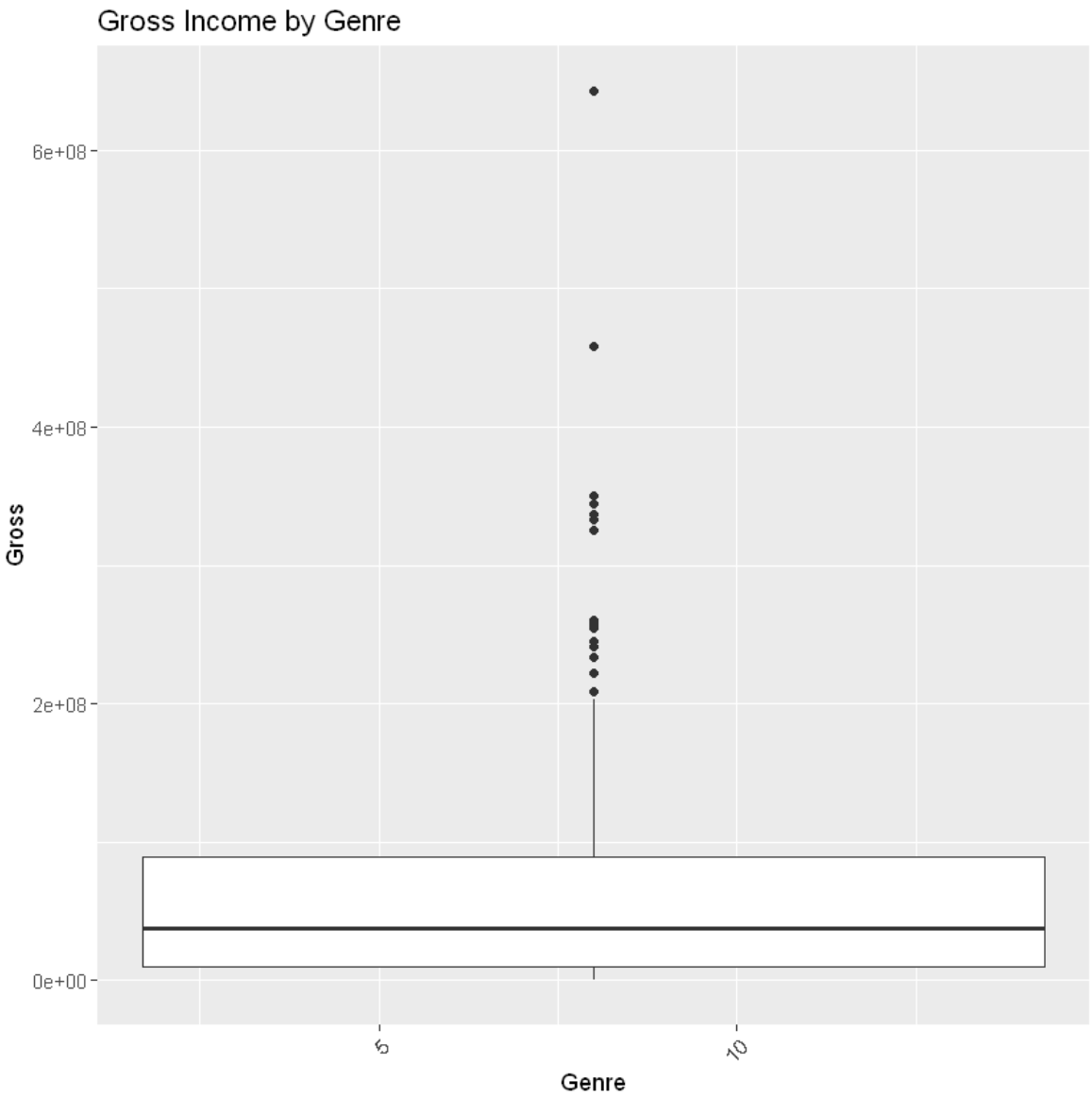


In [24]:
```
ggplot(data, aes(x = Ratings)) +
  geom_histogram(binwidth = 0.5, fill = "blue", color = "black") +
  labs(title = "Distribution of Movie Ratings",
       x = "Ratings",
       y = "Frequency")
```

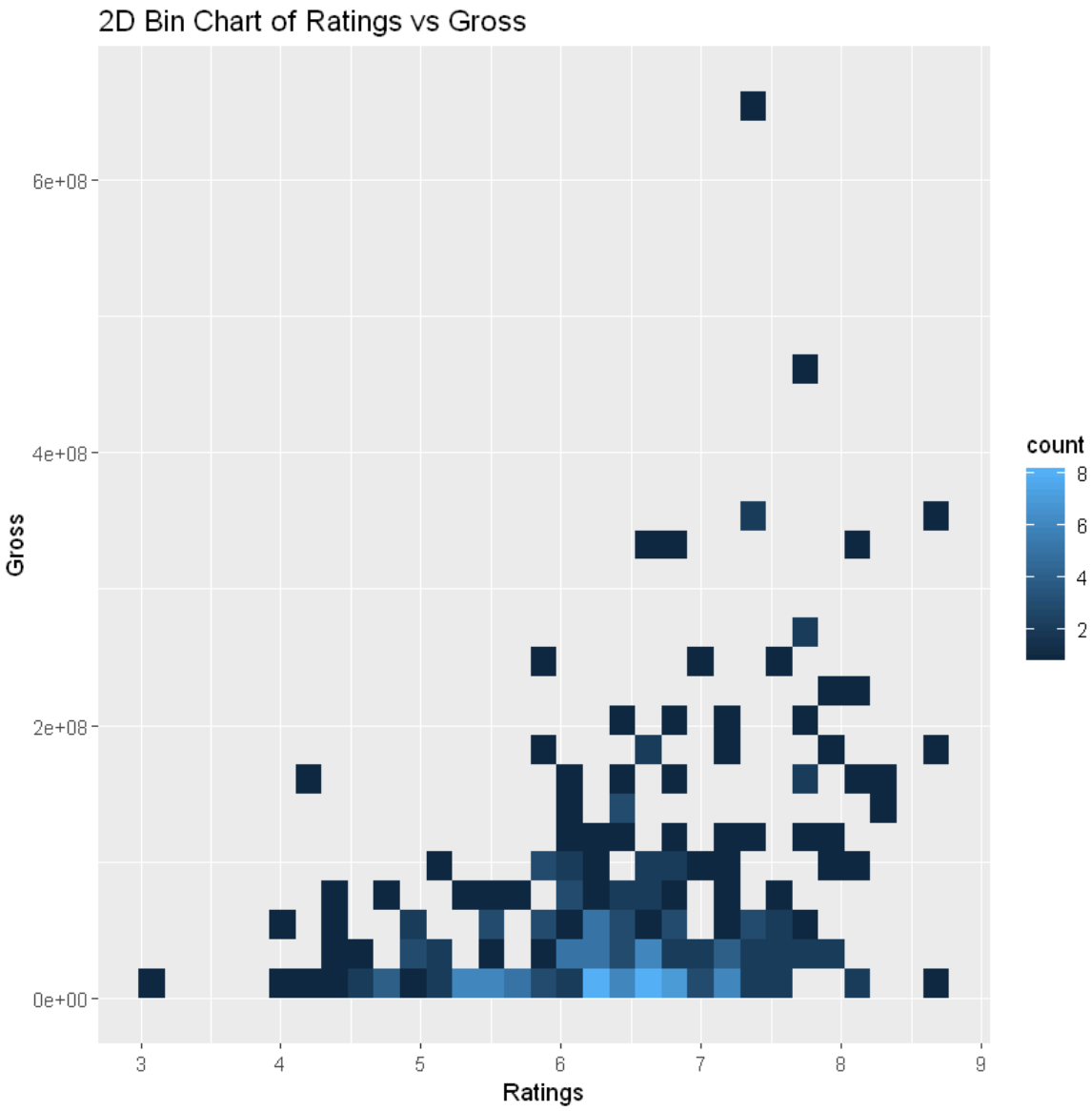


In [25]:
```
ggplot(data, aes(x = Genre, y = Gross)) +
  geom_boxplot() +
  labs(title = "Gross Income by Genre",
       x = "Genre",
       y = "Gross") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

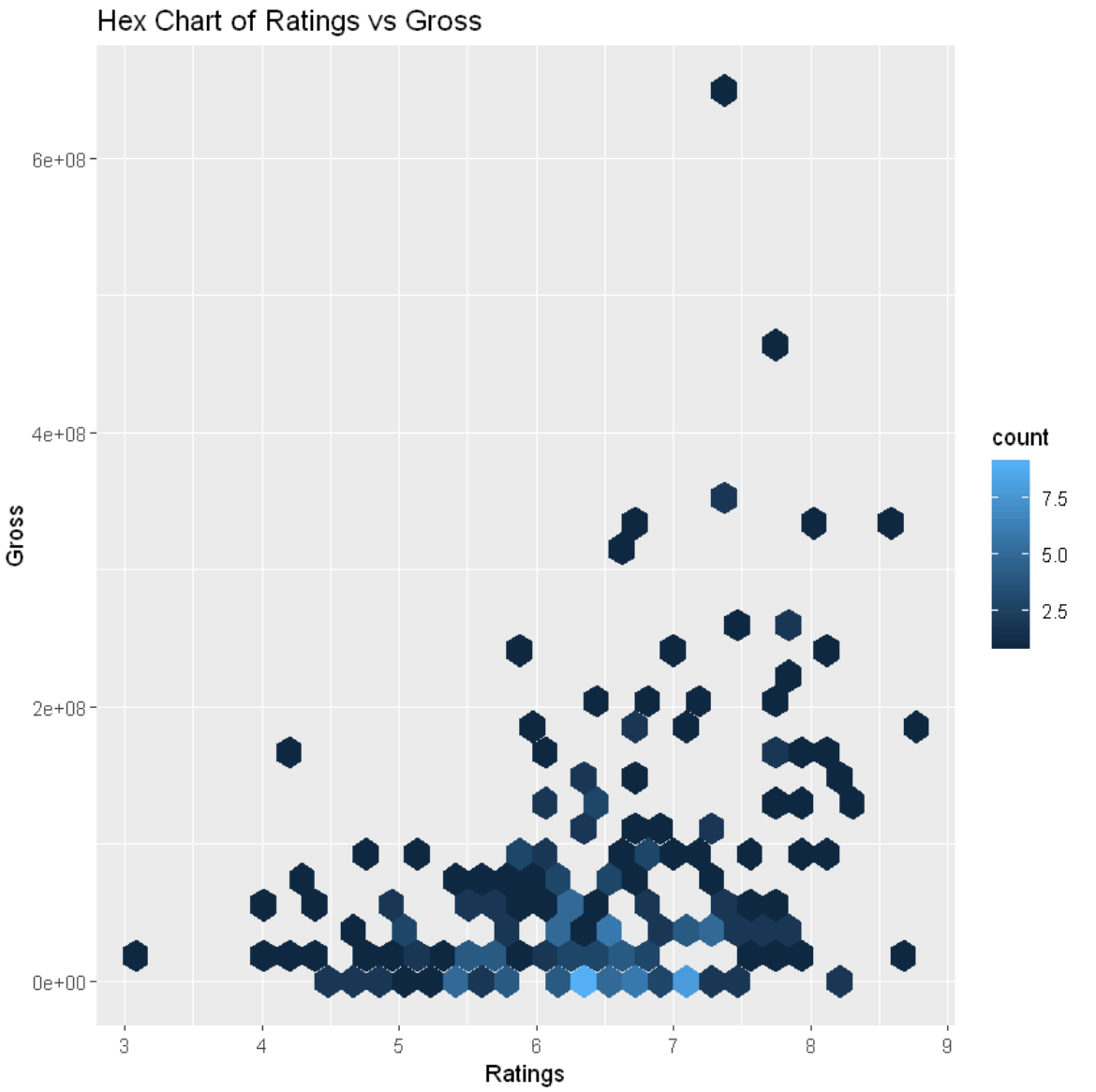
```
Warning message:
"Continuous x aesthetic -- did you forget aes(group=...)?"
```



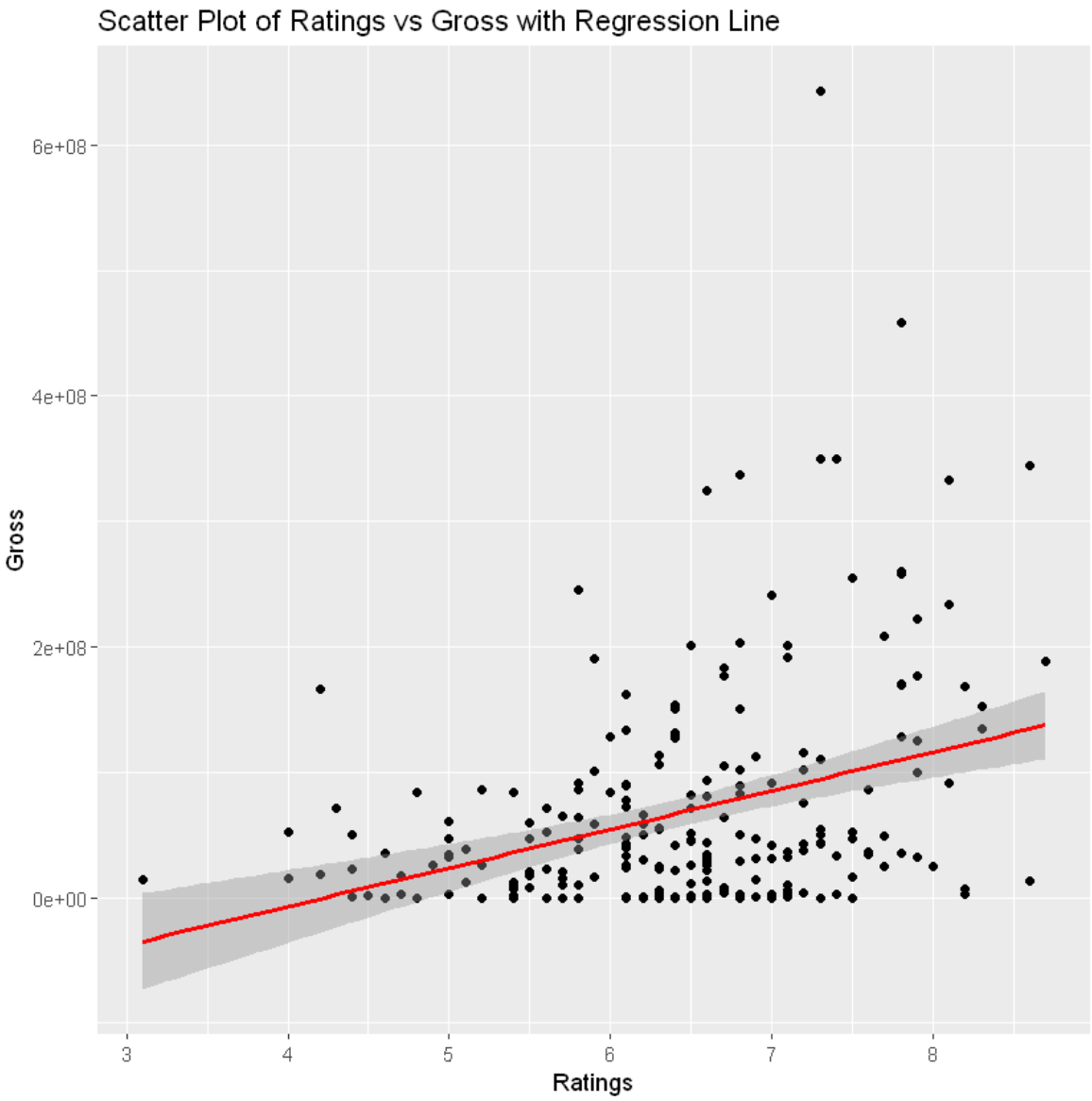
In [26]:
```
ggplot(data, aes(x = Ratings, y = Gross)) +
  geom_bin2d() +
  labs(title = "2D Bin Chart of Ratings vs Gross",
       x = "Ratings",
       y = "Gross")
```



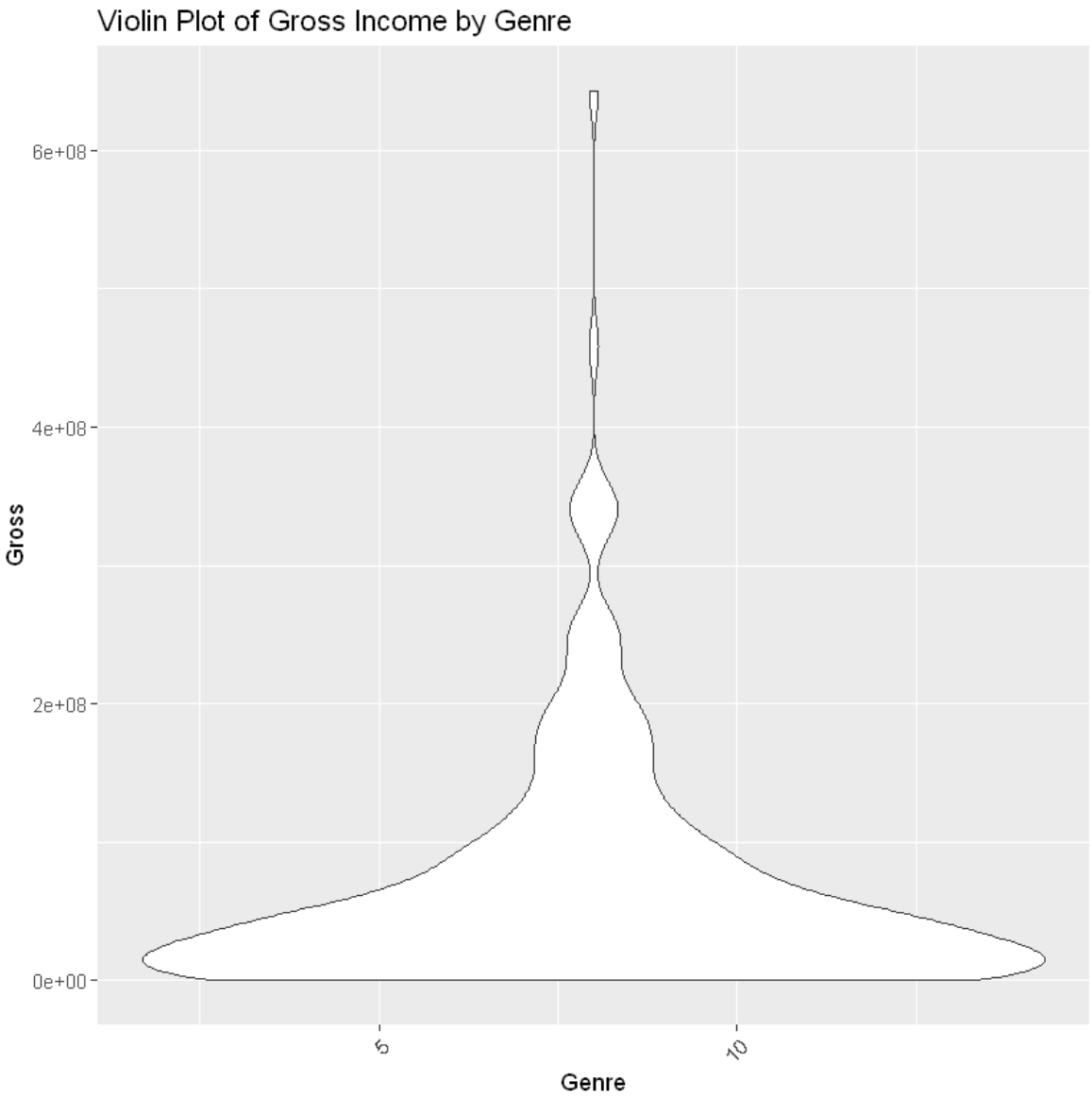
In [27]:
```
ggplot(data, aes(x = Ratings, y = Gross)) +
  geom_hex() +
  labs(title = "Hex Chart of Ratings vs Gross",
       x = "Ratings",
       y = "Gross")
```



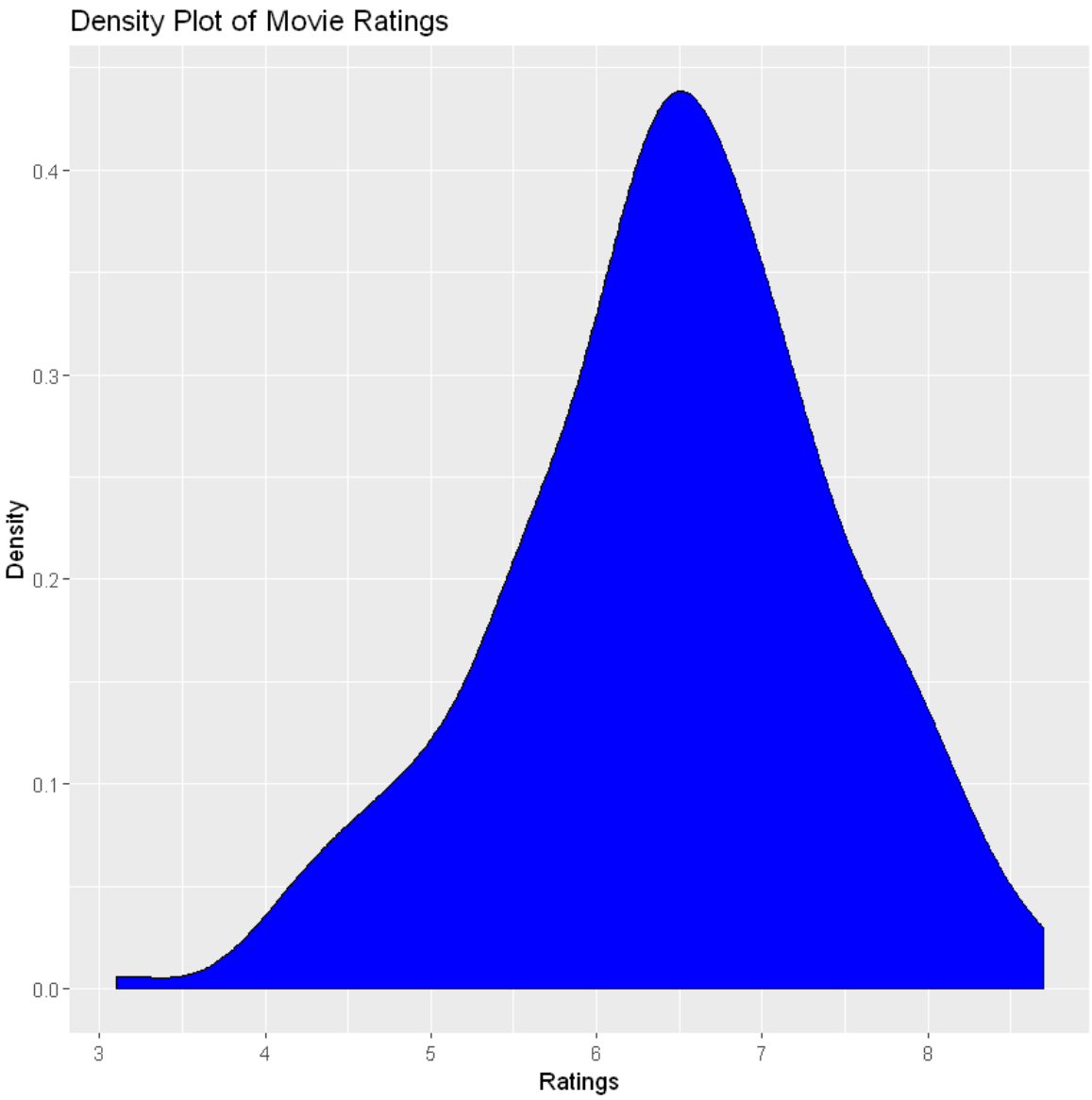
In [28]:
```
ggplot(data, aes(x = Ratings, y = Gross)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red") +
  labs(title = "Scatter Plot of Ratings vs Gross with Regression Line",
       x = "Ratings",
       y = "Gross")
```



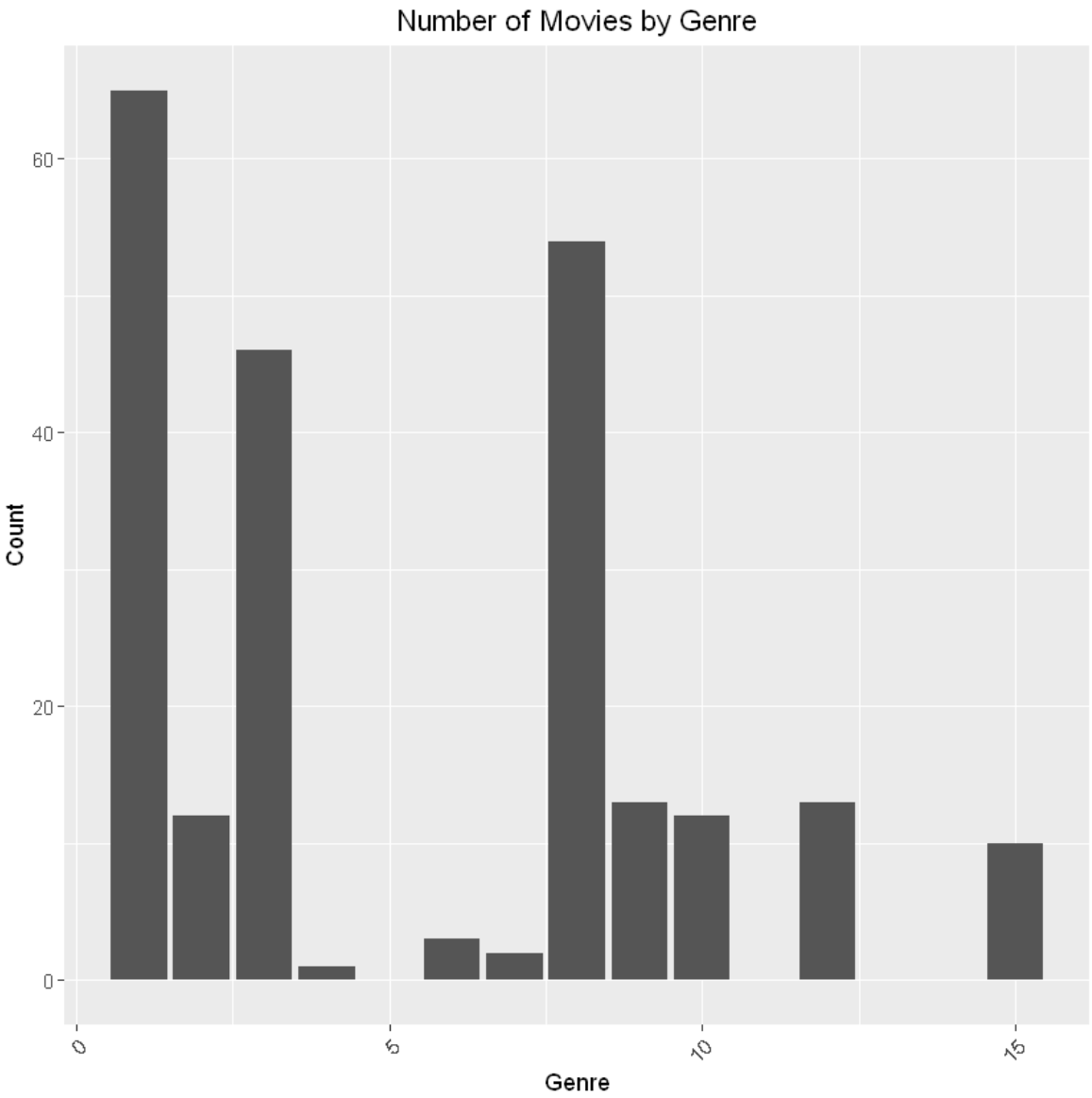
In [29]:
```
ggplot(data, aes(x = Genre, y = Gross)) +
  geom_violin() +
  labs(title = "Violin Plot of Gross Income by Genre",
       x = "Genre",
       y = "Gross") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



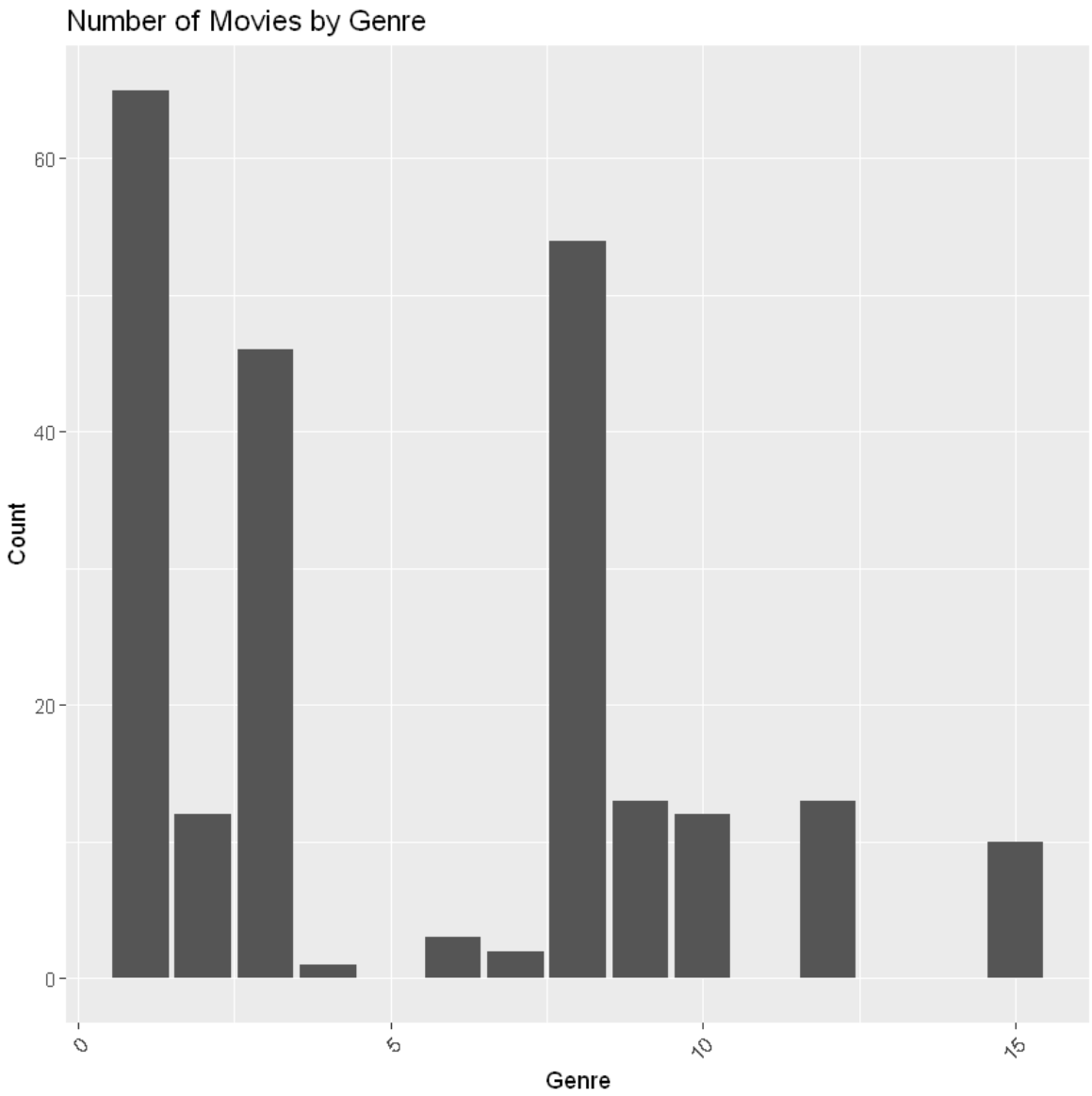
In [30]:
```
ggplot(data, aes(x = Ratings)) +
  geom_density(fill = "blue") +
  labs(title = "Density Plot of Movie Ratings",
       x = "Ratings",
       y = "Density")
```



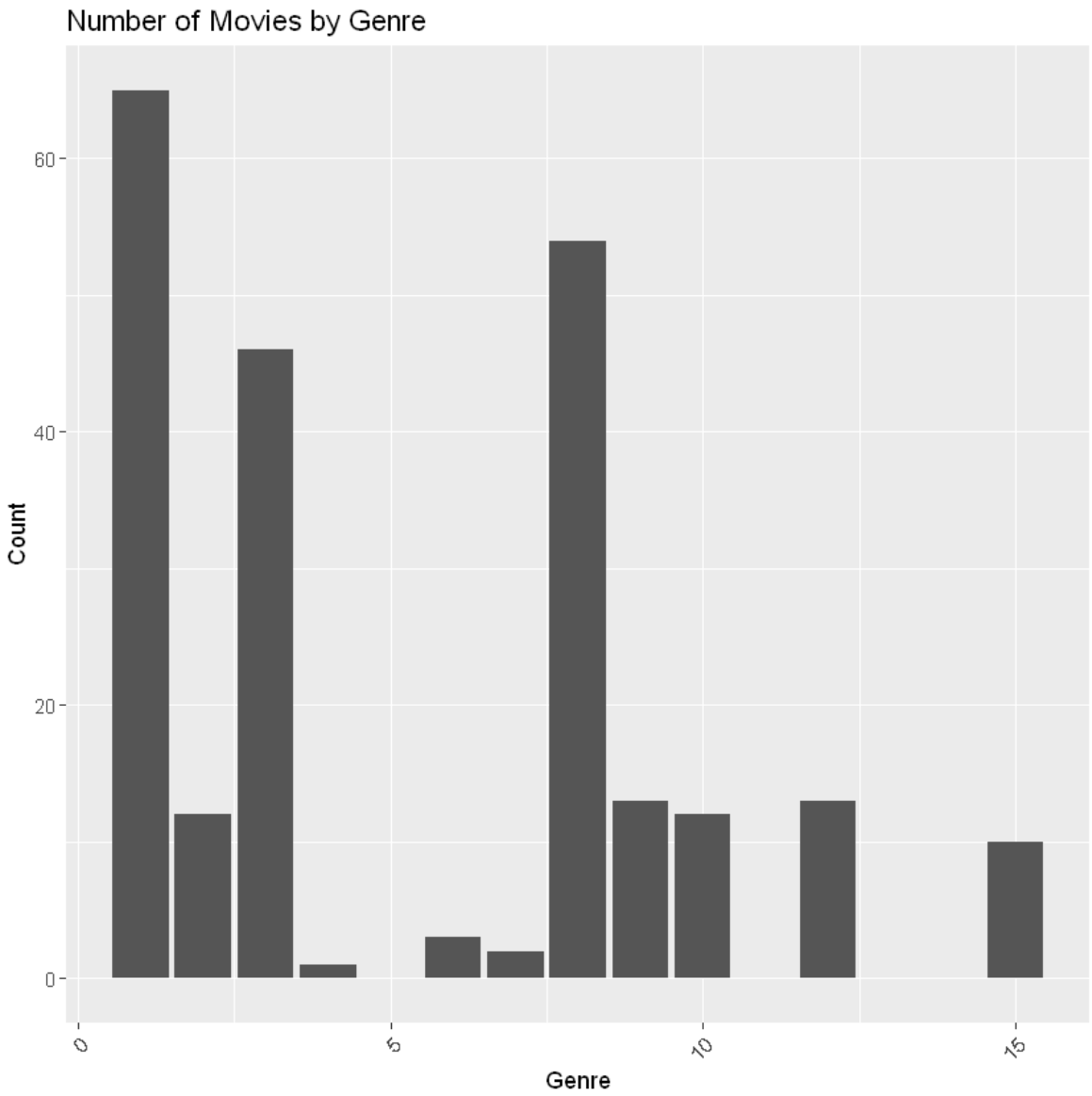
In [31]:
```
ggplot(data, aes(x = Genre, fill = Genre)) +
  geom_bar() +
  labs(title = "Number of Movies by Genre",
       x = "Genre",
       y = "Count",
       fill = "Genre Type") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```



In [32]:
```
ggplot(data, aes(x = Genre)) +
  geom_bar() +
  labs(title = "Number of Movies by Genre",
       x = "Genre",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

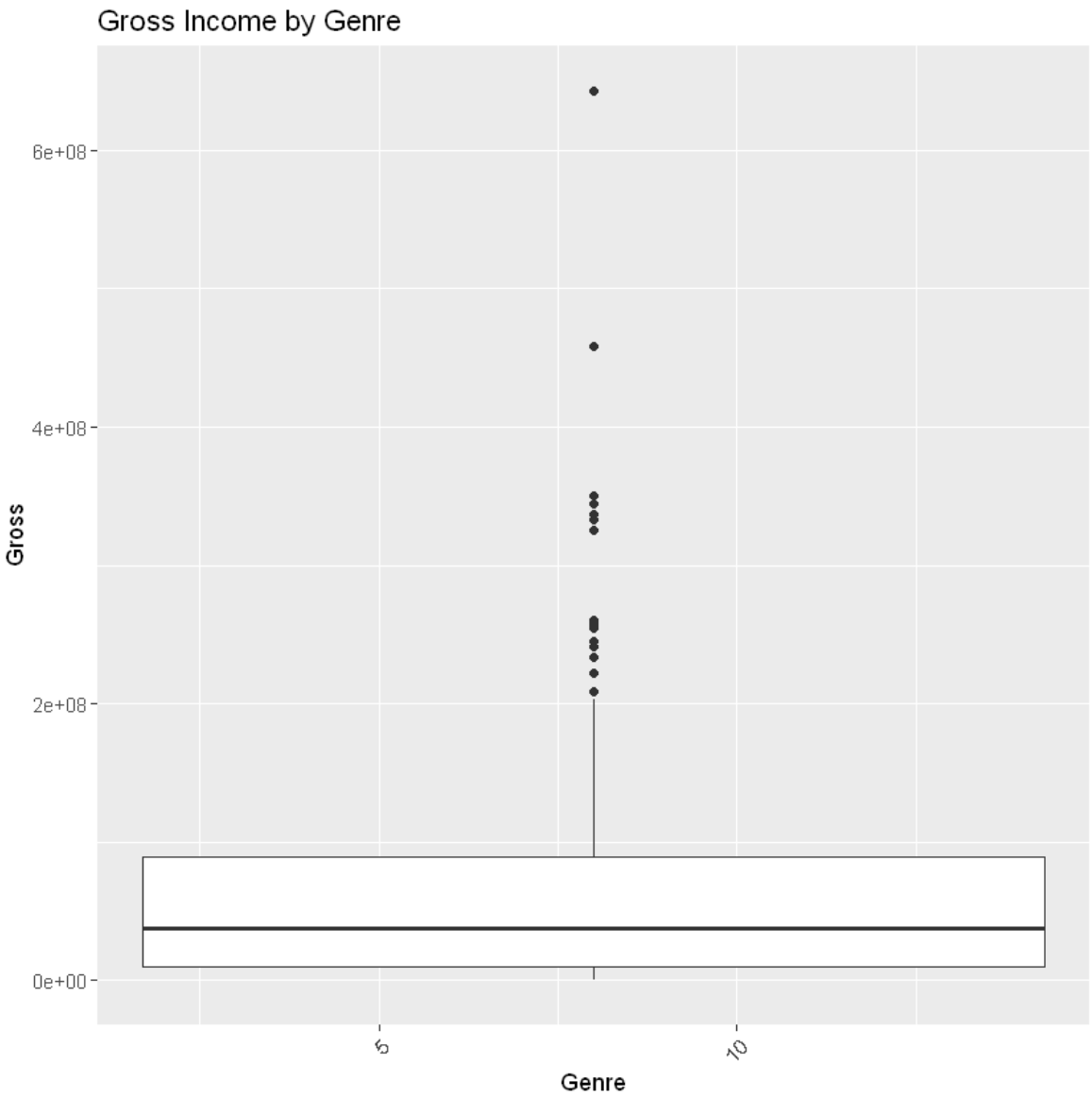


In [33]:
```
ggplot(data, aes(x = Genre)) +
  geom_bar() +
  labs(title = "Number of Movies by Genre",
       x = "Genre",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

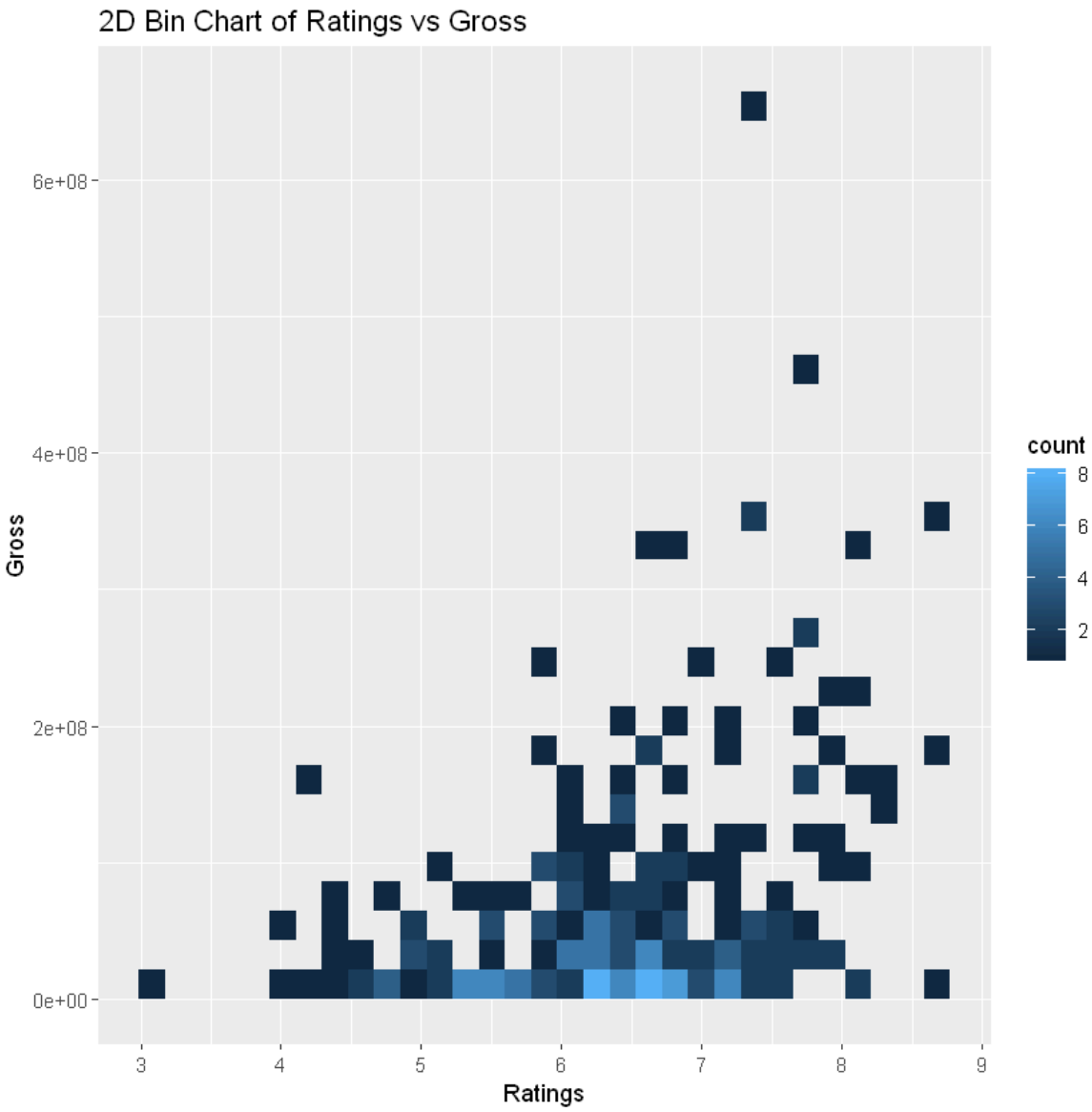


In [34]:
```
ggplot(data, aes(x = Genre, y = Gross)) +
  geom_boxplot() +
  labs(title = "Gross Income by Genre",
       x = "Genre",
       y = "Gross") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

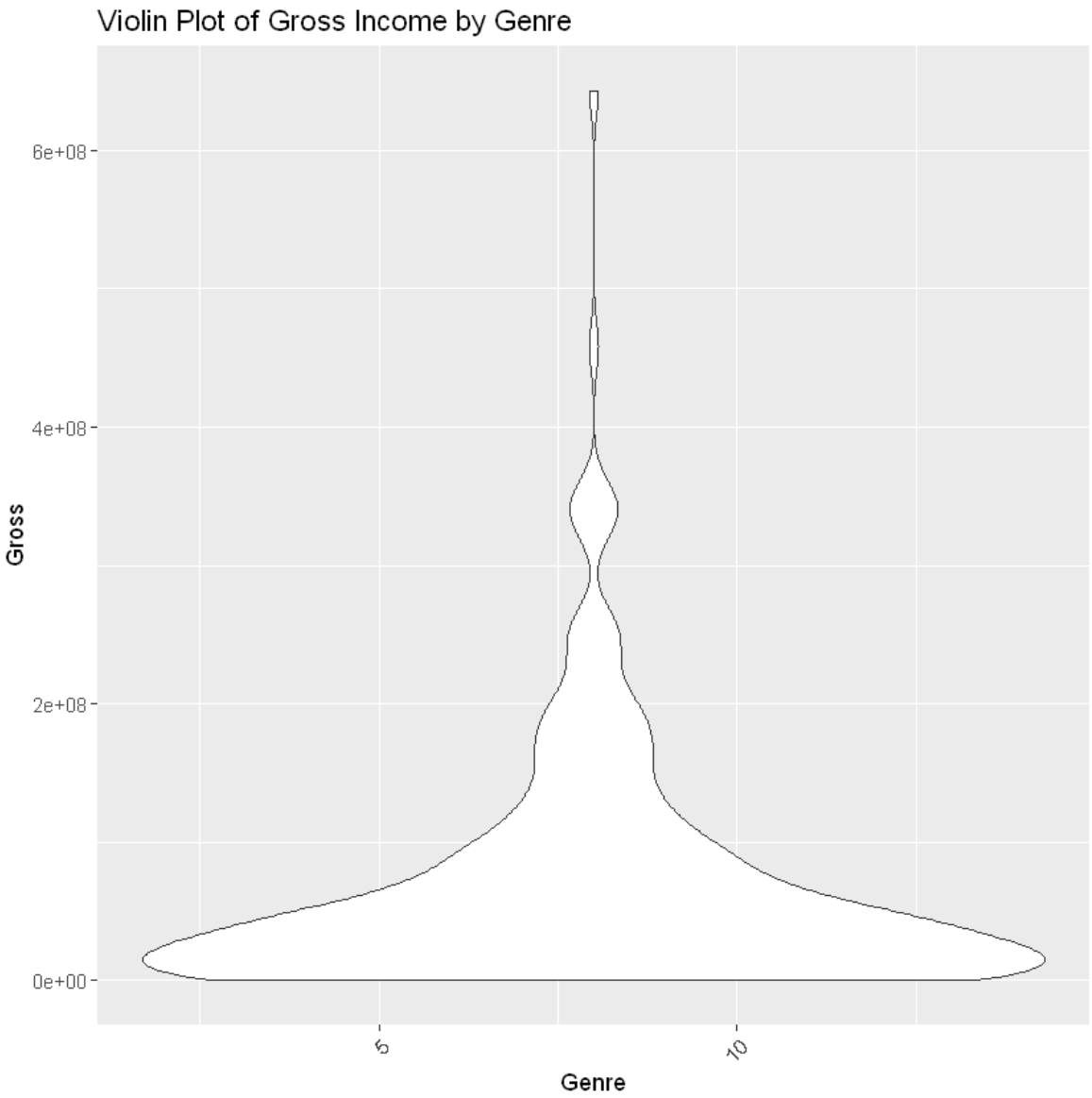
```
Warning message:
"Continuous x aesthetic -- did you forget aes(group=...)?"
```



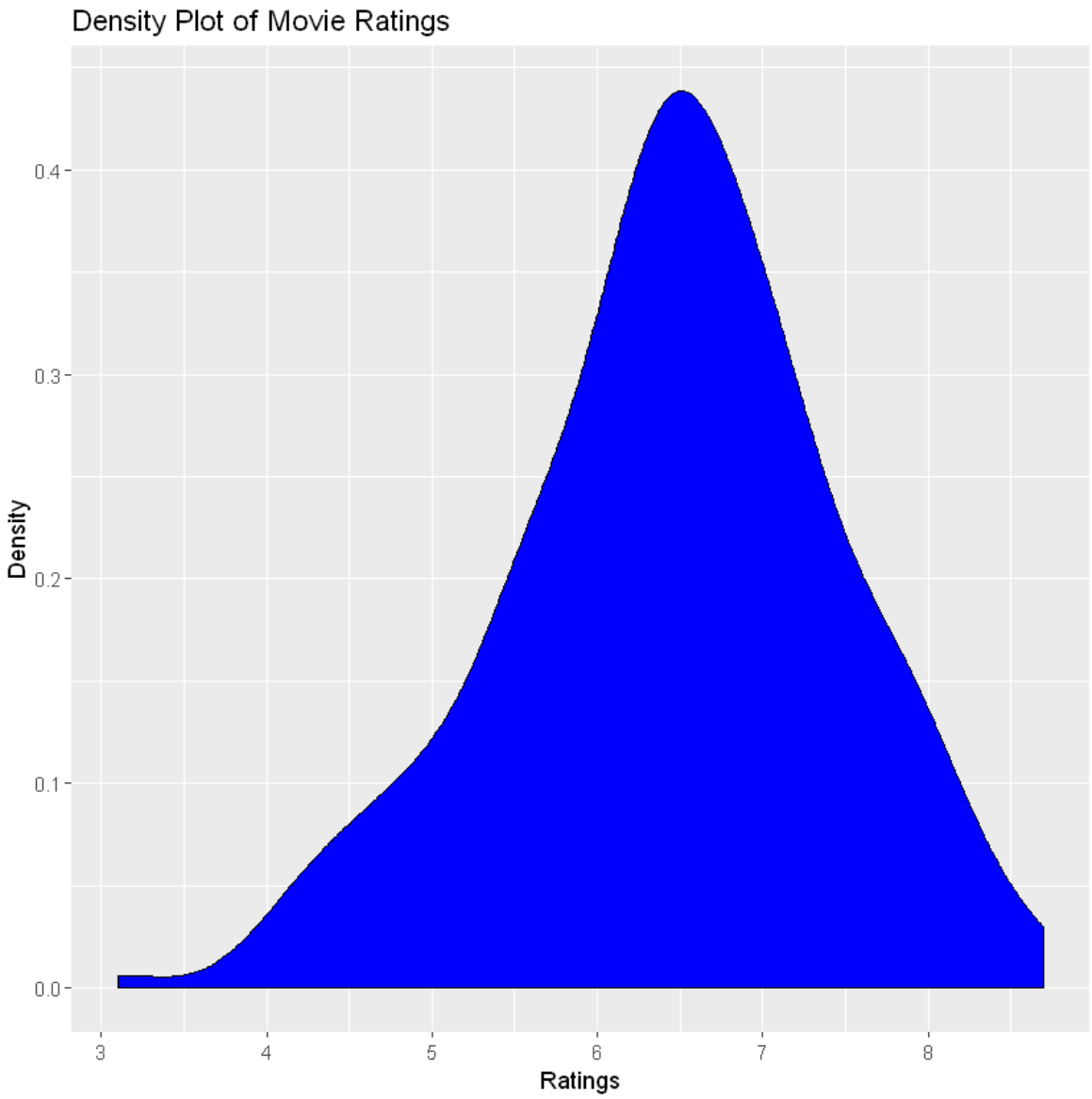
In [35]:
```
ggplot(data, aes(x = Ratings, y = Gross)) +
  geom_bin2d() +
  labs(title = "2D Bin Chart of Ratings vs Gross",
       x = "Ratings",
       y = "Gross")
```



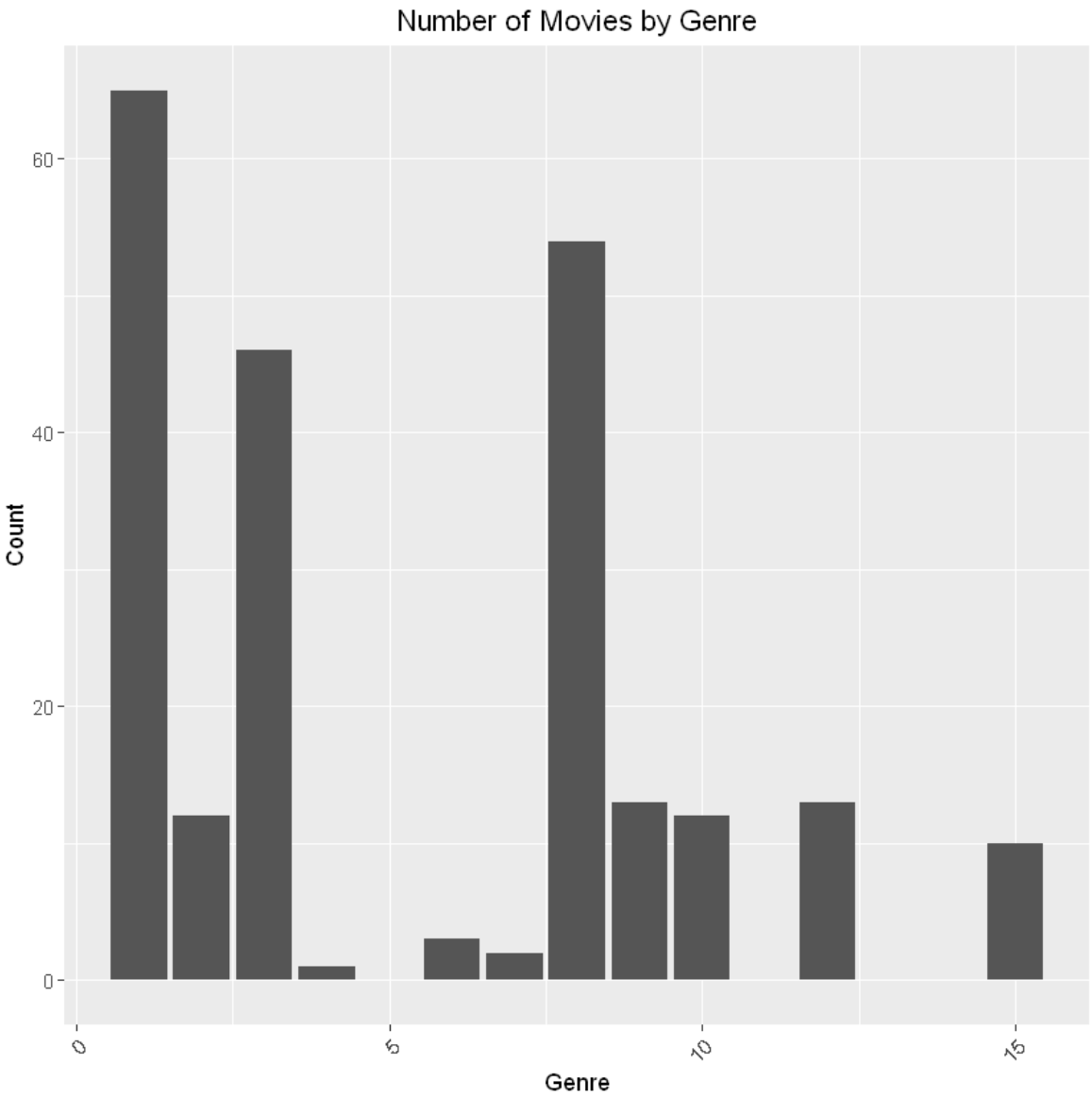
In [36]:
```
ggplot(data, aes(x = Genre, y = Gross)) +
  geom_violin() +
  labs(title = "Violin Plot of Gross Income by Genre",
       x = "Genre",
       y = "Gross") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



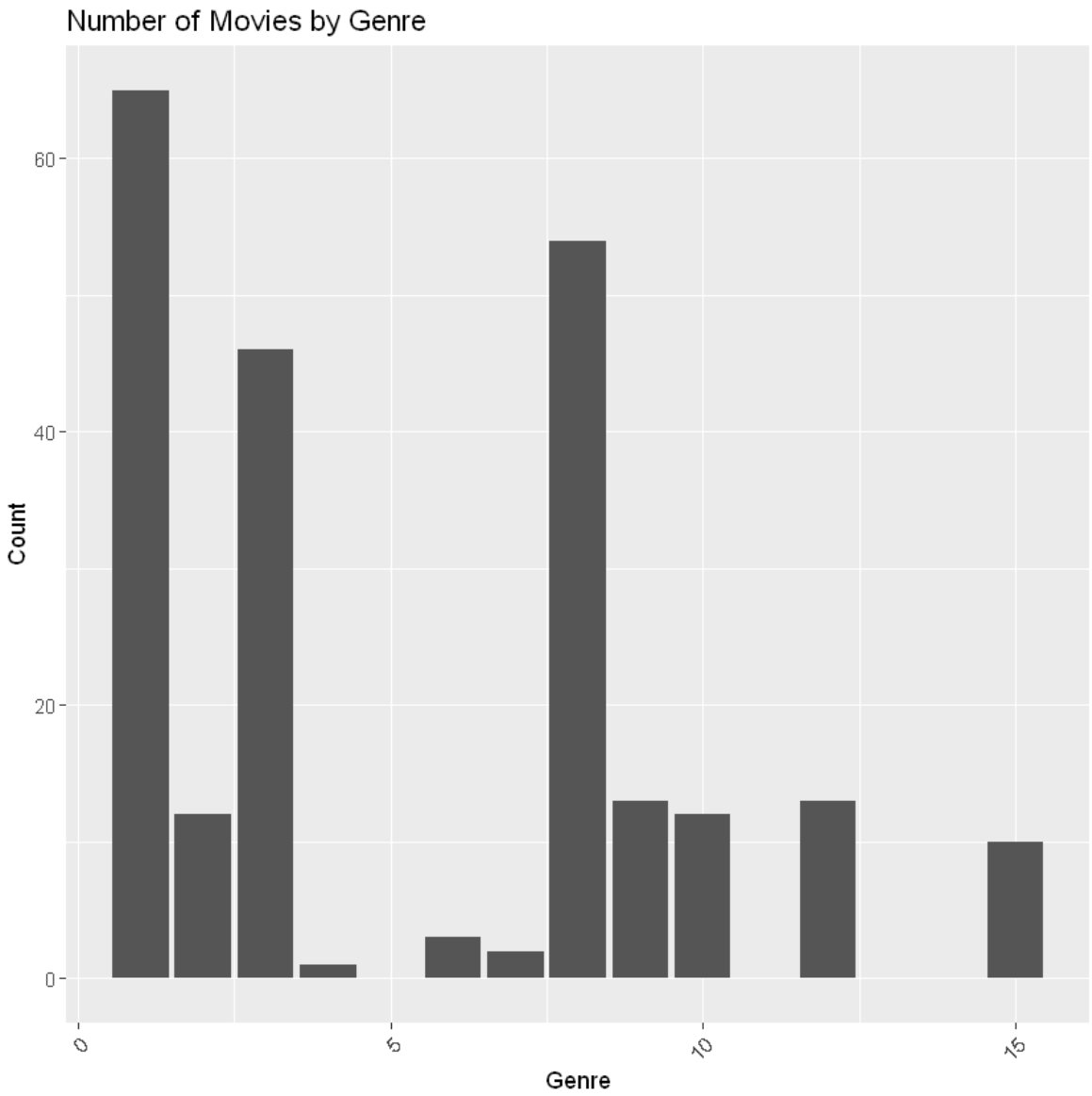
In [37]:
```
ggplot(data, aes(x = Ratings)) +
  geom_density(fill = "blue") +
  labs(title = "Density Plot of Movie Ratings",
       x = "Ratings",
       y = "Density")
```



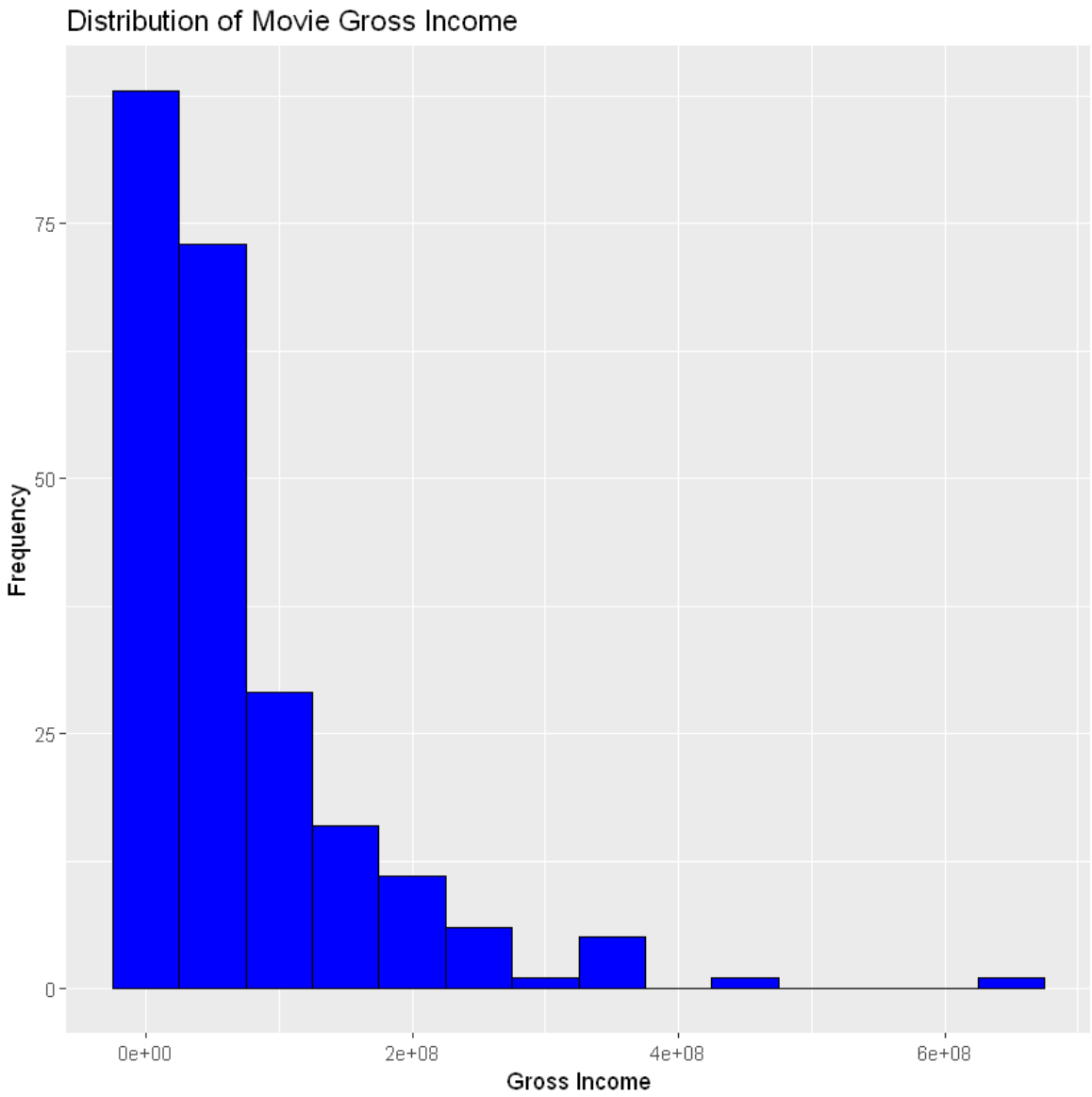
In [38]:
```
ggplot(data, aes(x = Genre, fill = Genre)) +
  geom_bar() +
  labs(title = "Number of Movies by Genre",
       x = "Genre",
       y = "Count",
       fill = "Genre Type") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```



In [39]:
```
ggplot(data, aes(x = Genre)) +
  geom_bar() +
  labs(title = "Number of Movies by Genre",
       x = "Genre",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

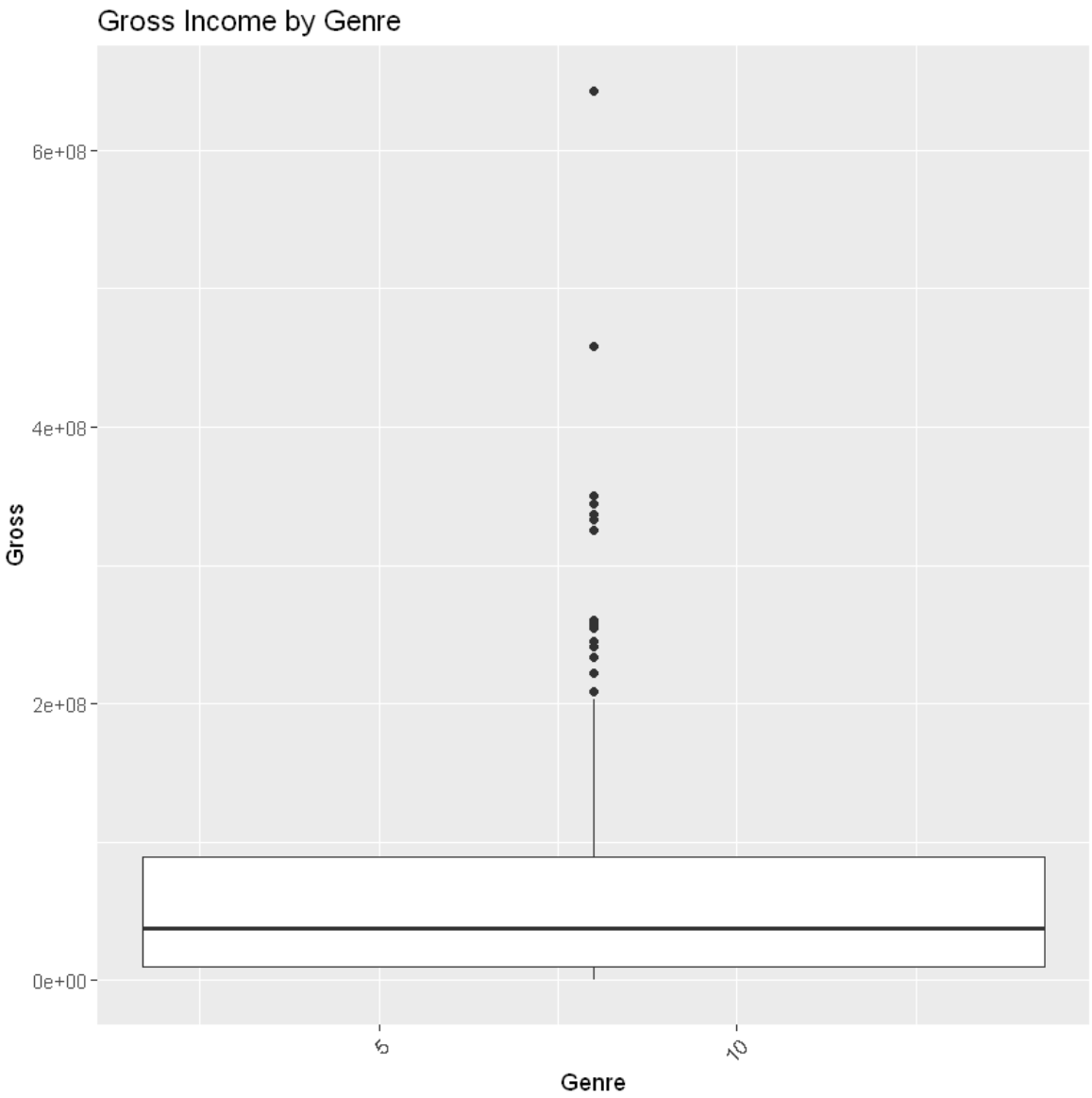


In [40]:
```
ggplot(data, aes(x = Gross)) +
  geom_histogram(binwidth = 50000000, fill = "blue", color = "black") +
  labs(title = "Distribution of Movie Gross Income",
       x = "Gross Income",
       y = "Frequency")
```

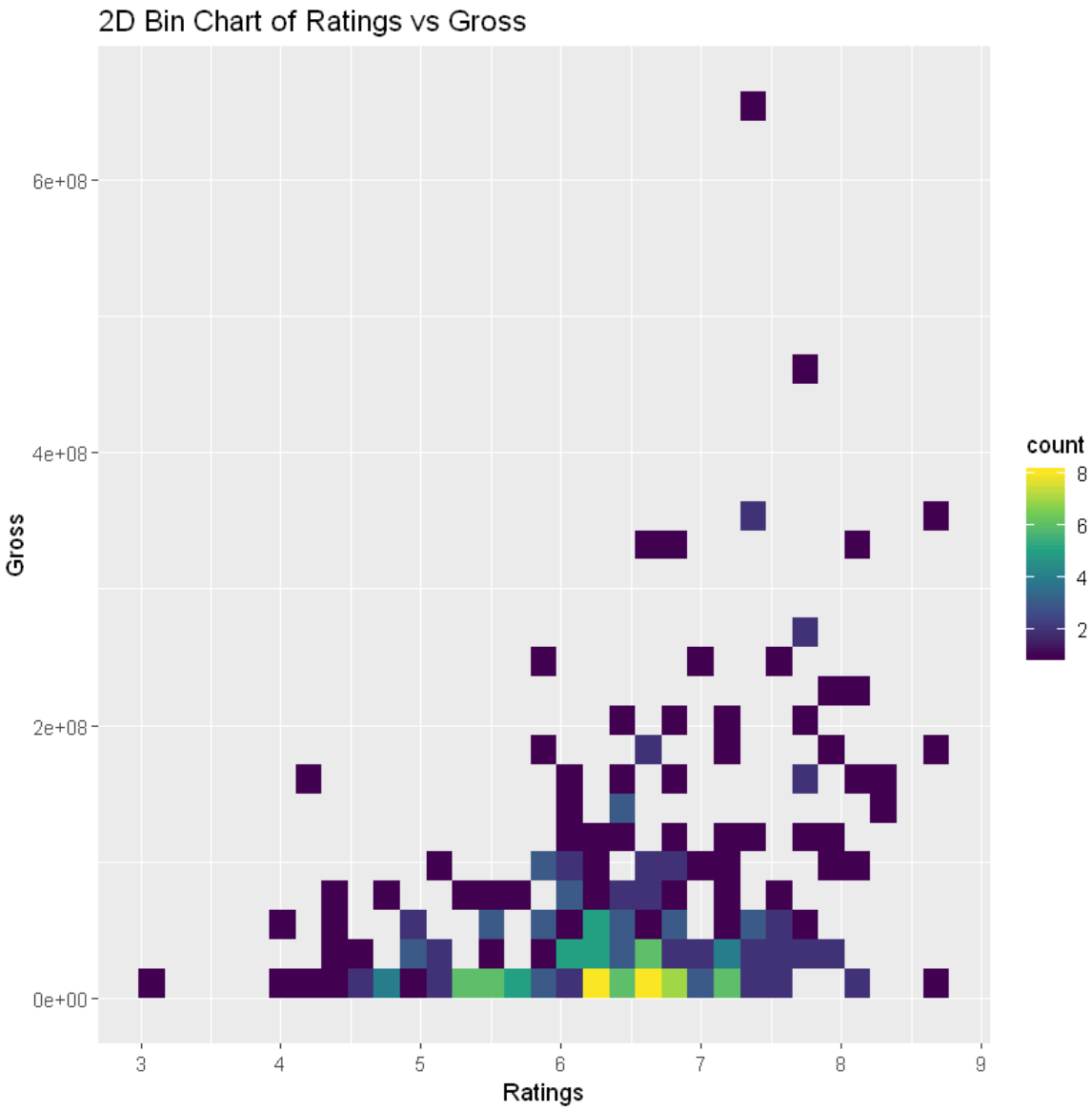


In [41]:
```
ggplot(data, aes(x = Genre, y = Gross)) +
  geom_boxplot() +
  labs(title = "Gross Income by Genre",
       x = "Genre",
       y = "Gross") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

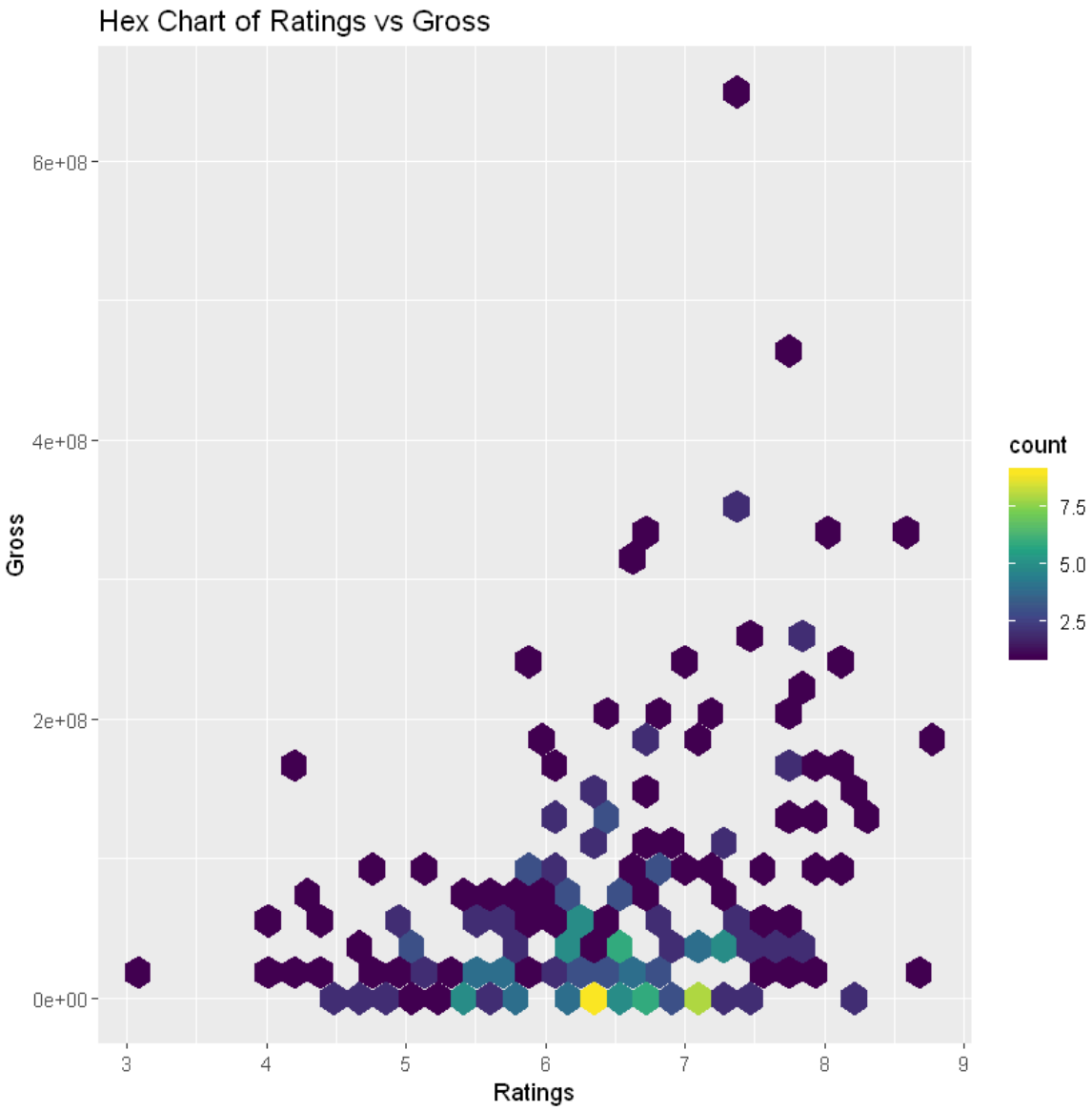
```
Warning message:
"Continuous x aesthetic -- did you forget aes(group=...)?"
```



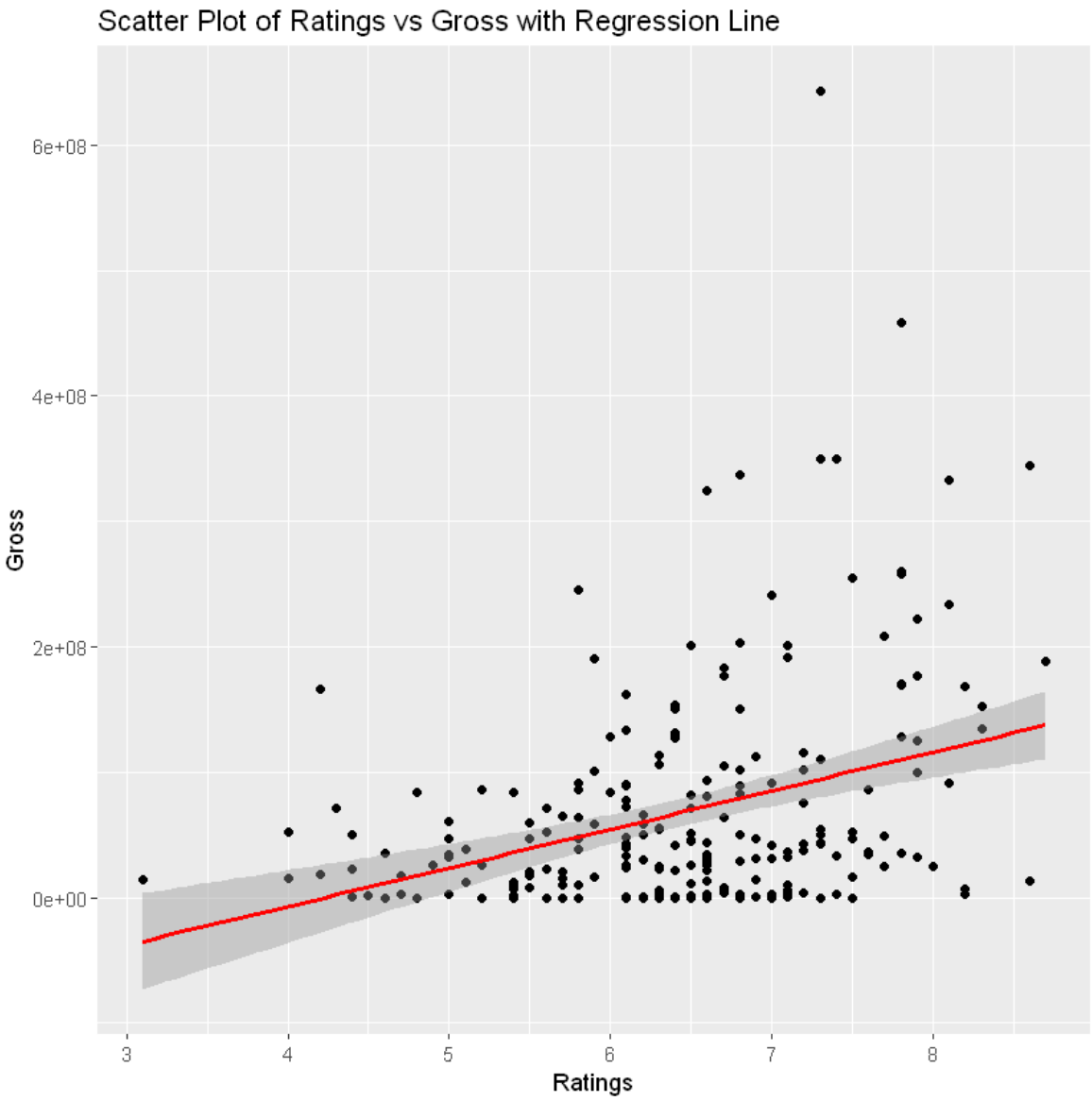
In [42]:
```
ggplot(data, aes(x = Ratings, y = Gross)) +
  geom_bin2d() +
  labs(title = "2D Bin Chart of Ratings vs Gross",
       x = "Ratings",
       y = "Gross") +
  scale_fill_continuous(type = "viridis")
```



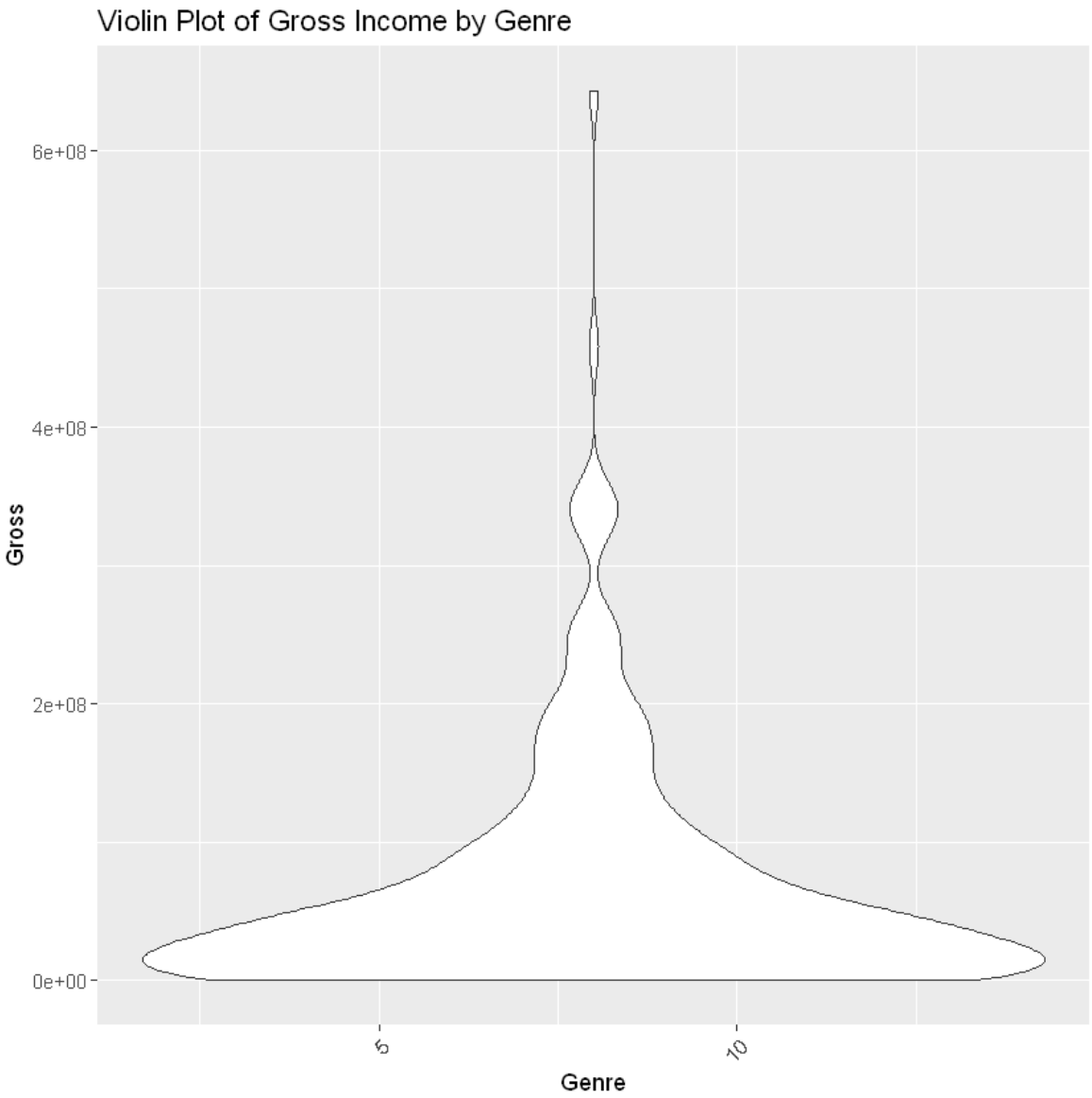
In [43]:
```
ggplot(data, aes(x = Ratings, y = Gross)) +
  geom_hex() +
  labs(title = "Hex Chart of Ratings vs Gross",
       x = "Ratings",
       y = "Gross") +
  scale_fill_continuous(type = "viridis")
```



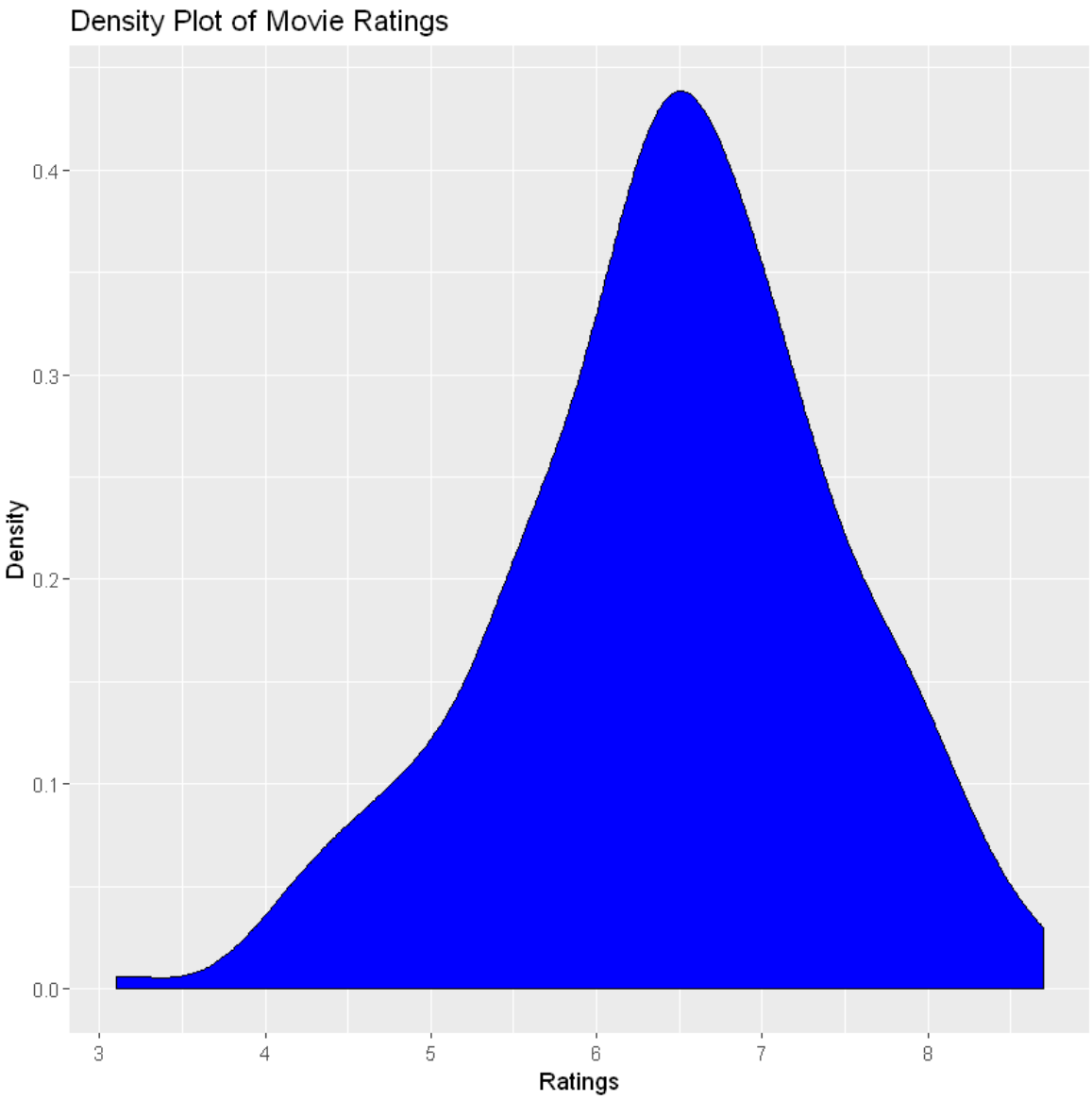
In [44]:
```
ggplot(data, aes(x = Ratings, y = Gross)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red") +
  labs(title = "Scatter Plot of Ratings vs Gross with Regression Line",
       x = "Ratings",
       y = "Gross")
```



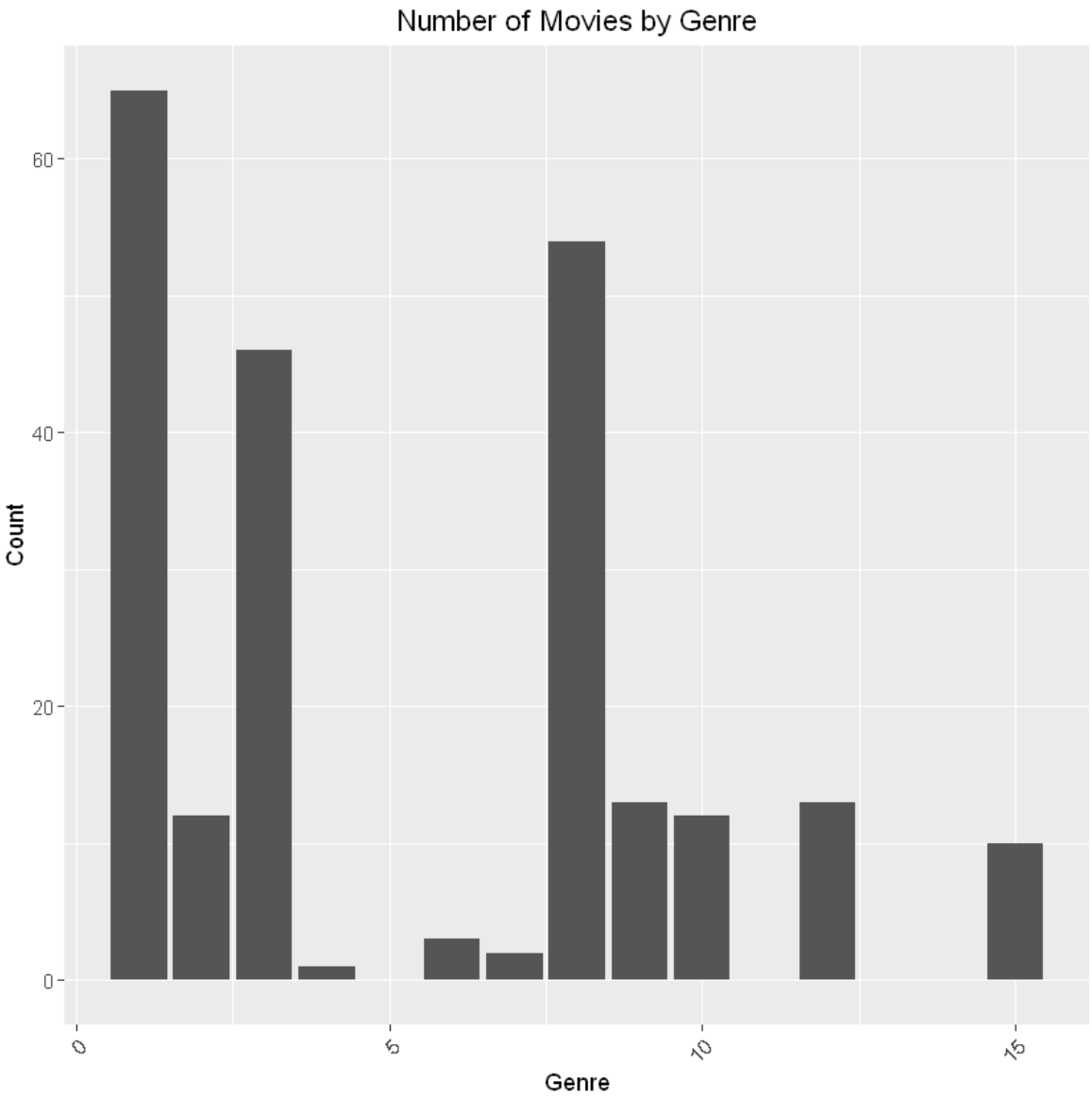
In [45]:
```
ggplot(data, aes(x = Genre, y = Gross)) +
  geom_violin() +
  labs(title = "Violin Plot of Gross Income by Genre",
       x = "Genre",
       y = "Gross") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



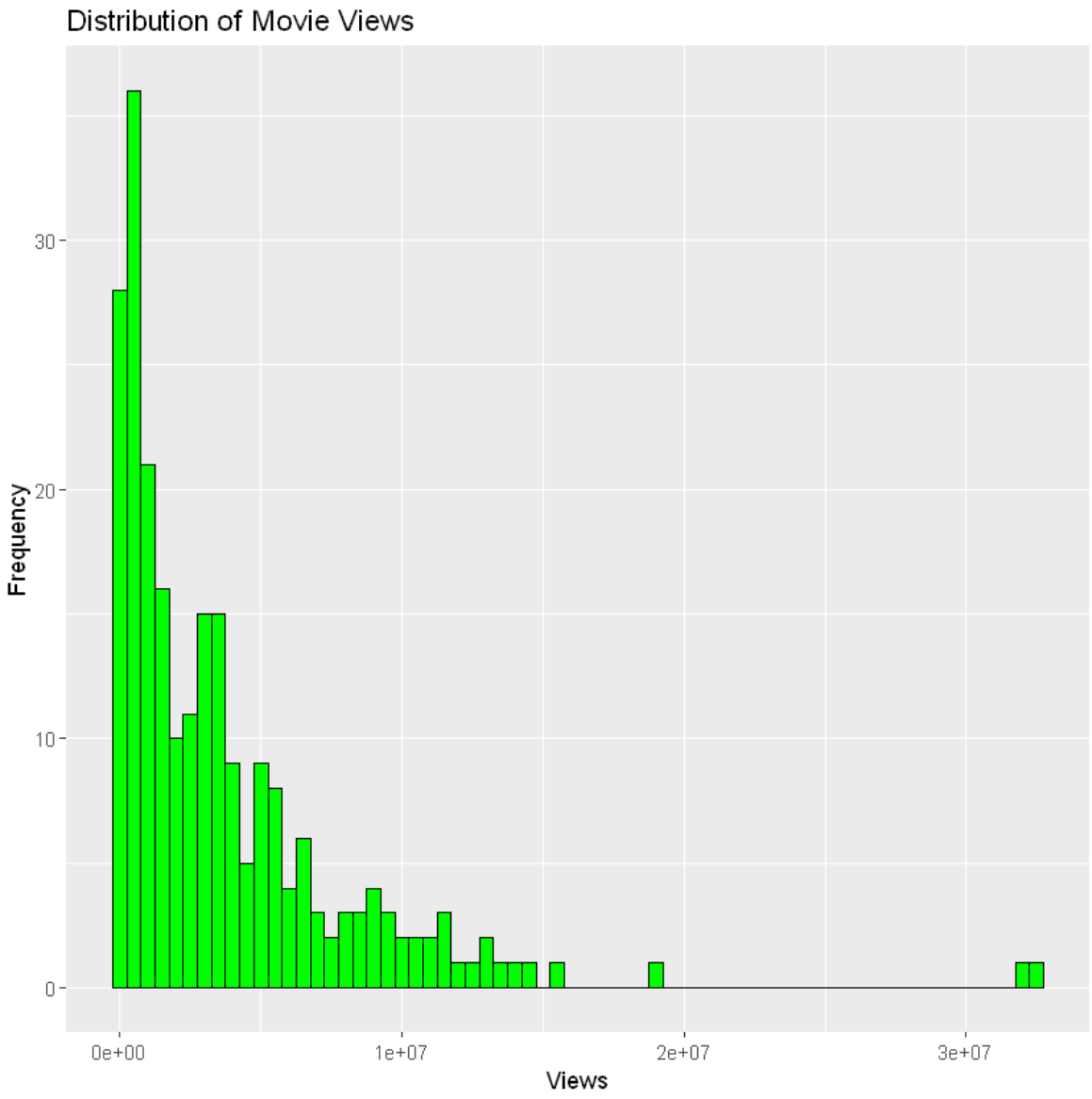
In [46]:
```
ggplot(data, aes(x = Ratings)) +
  geom_density(fill = "blue") +
  labs(title = "Density Plot of Movie Ratings",
       x = "Ratings",
       y = "Density")
```



In [47]:
```
ggplot(data, aes(x = Genre, fill = Genre)) +
  geom_bar() +
  labs(title = "Number of Movies by Genre",
       x = "Genre",
       y = "Count",
       fill = "Genre Type") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```

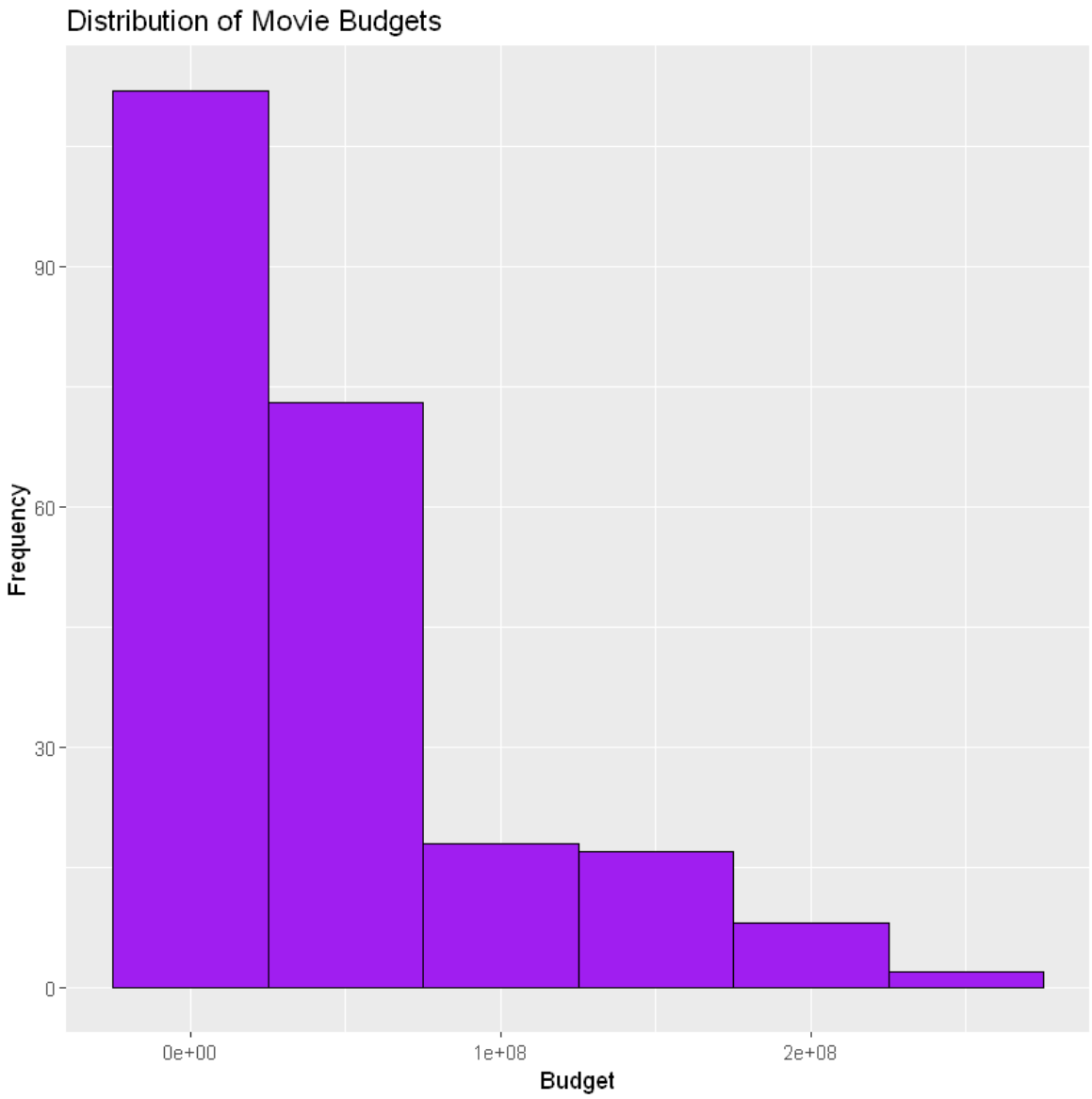


In [48]:
```
ggplot(data, aes(x = Views)) +
  geom_histogram(binwidth = 500000, fill = "green", color = "black") +
  labs(title = "Distribution of Movie Views",
       x = "Views",
       y = "Frequency")
```

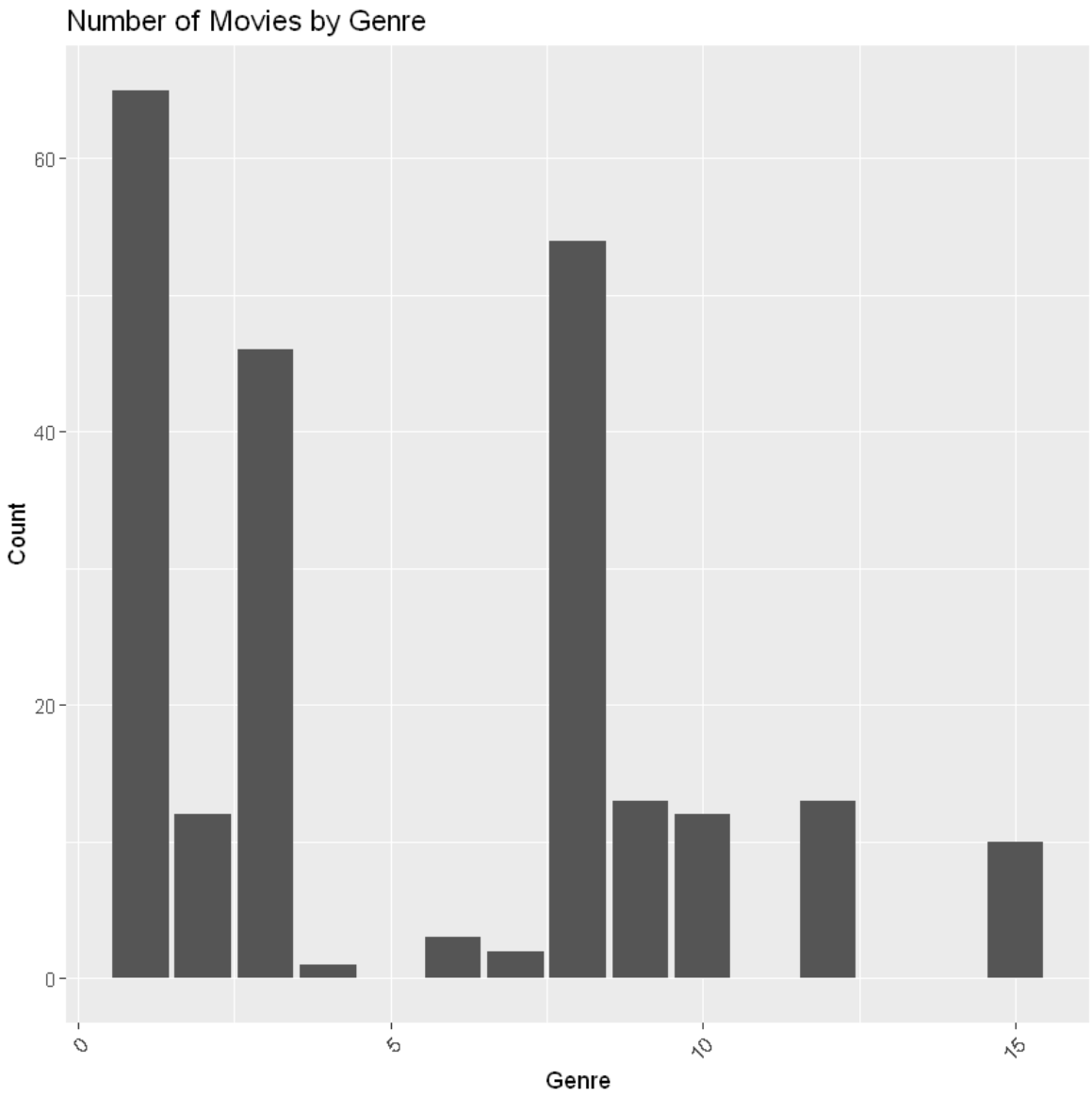


In [49]:
```
ggplot(data, aes(x = Budget)) +
  geom_histogram(binwidth = 50000000, fill = "purple", color = "black") +
  labs(title = "Distribution of Movie Budgets",
       x = "Budget",
       y = "Frequency")
```

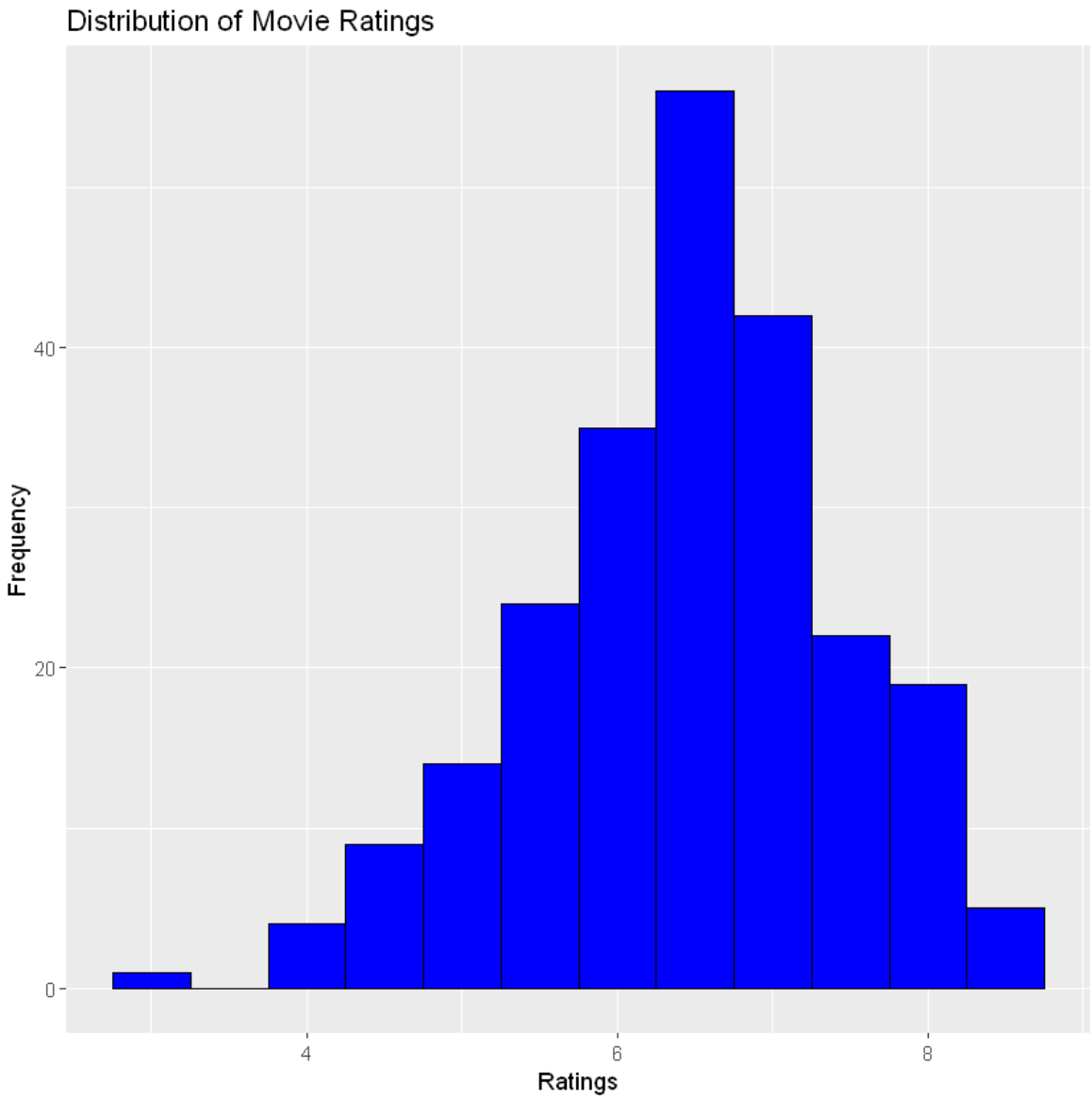
```
Warning message:
"Removed 1 rows containing non-finite values (stat_bin)."
```



In [50]:
```
ggplot(data, aes(x = Genre)) +
  geom_bar() +
  labs(title = "Number of Movies by Genre",
       x = "Genre",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

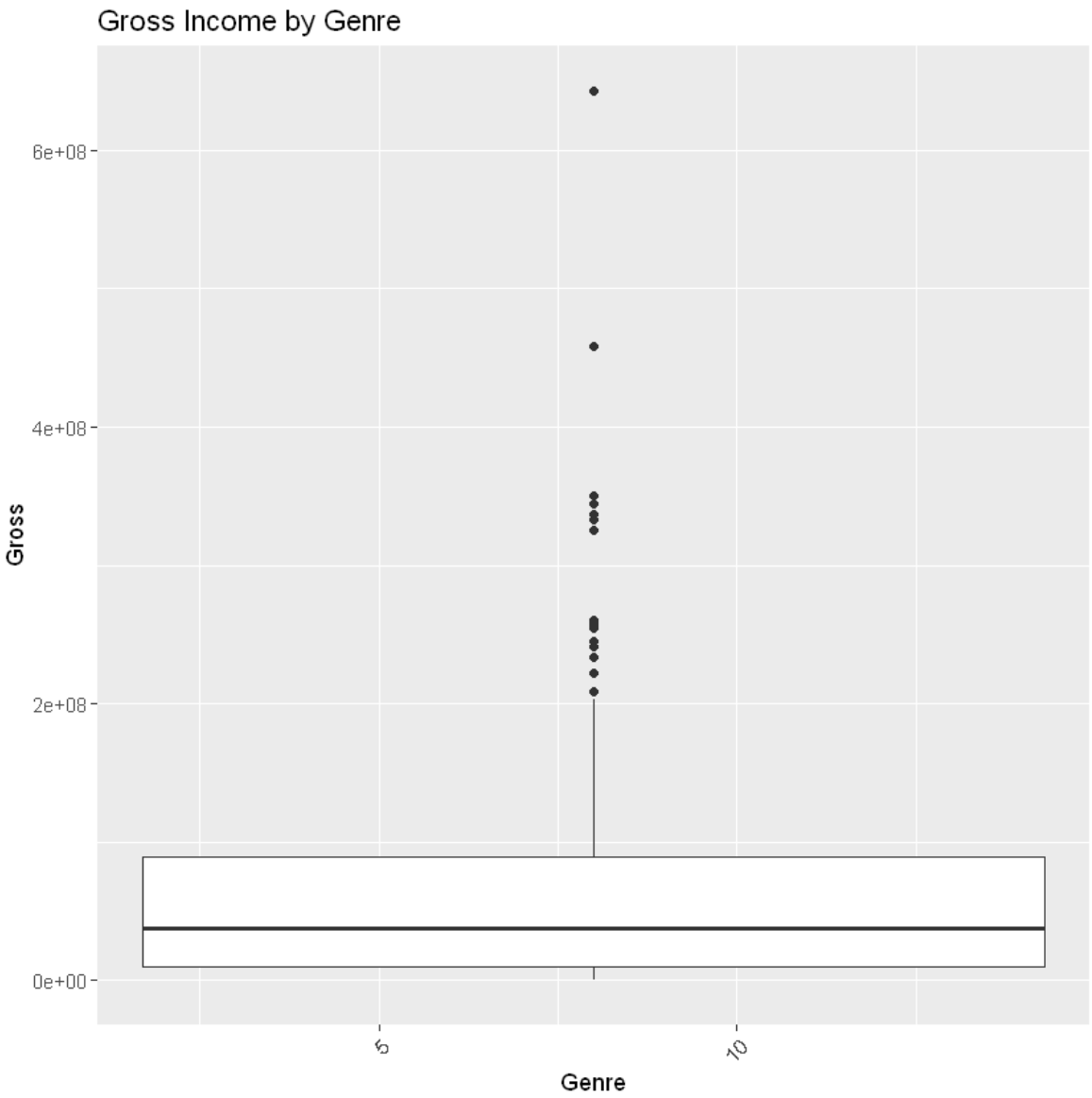


In [51]:
```
ggplot(data, aes(x = Ratings)) +
  geom_histogram(binwidth = 0.5, fill = "blue", color = "black") +
  labs(title = "Distribution of Movie Ratings",
       x = "Ratings",
       y = "Frequency")
```



In [52]:
```
ggplot(data, aes(x = Genre, y = Gross)) +
  geom_boxplot() +
  labs(title = "Gross Income by Genre",
       x = "Genre",
       y = "Gross") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

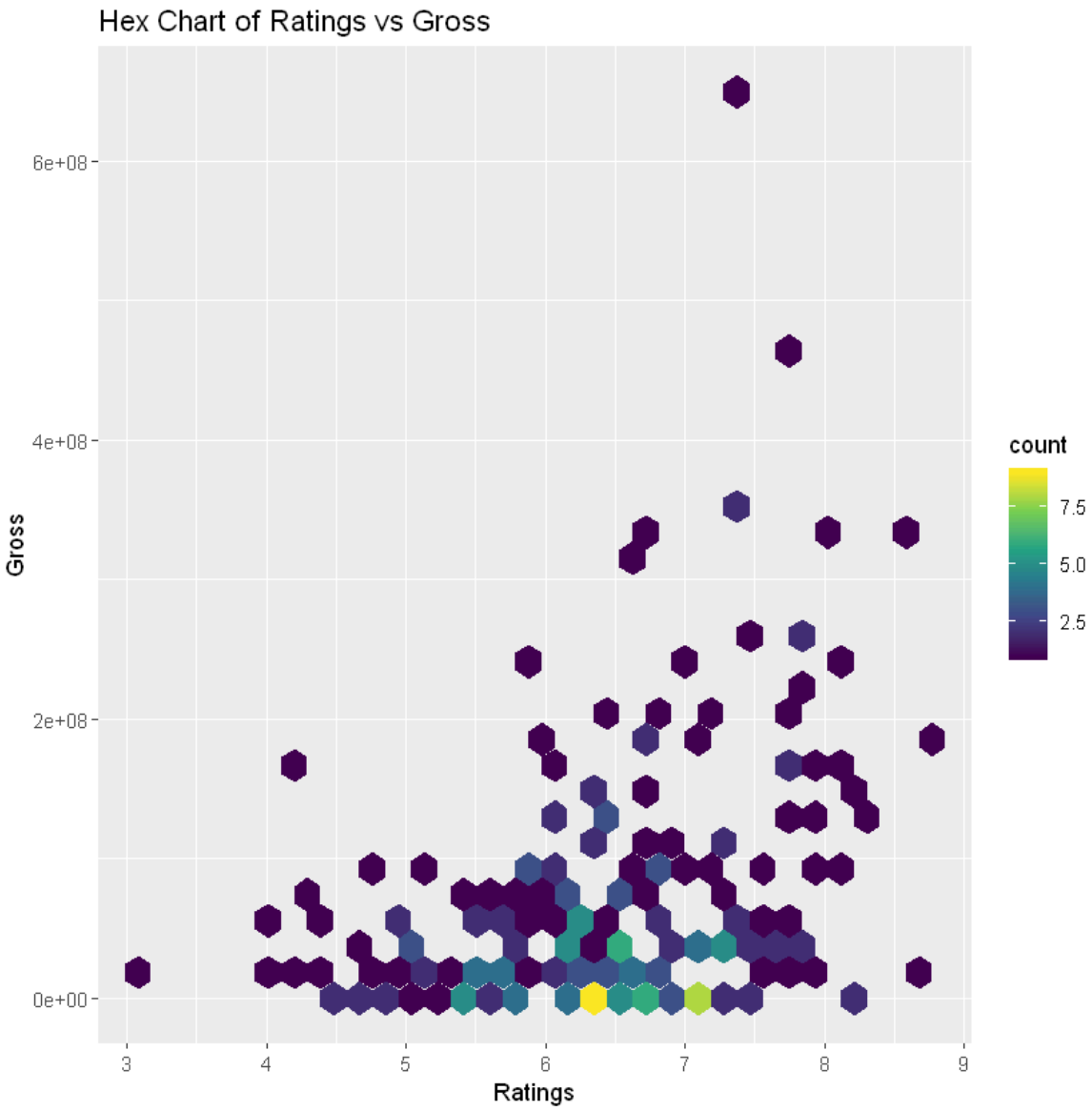
```
Warning message:
"Continuous x aesthetic -- did you forget aes(group=...)?"
```



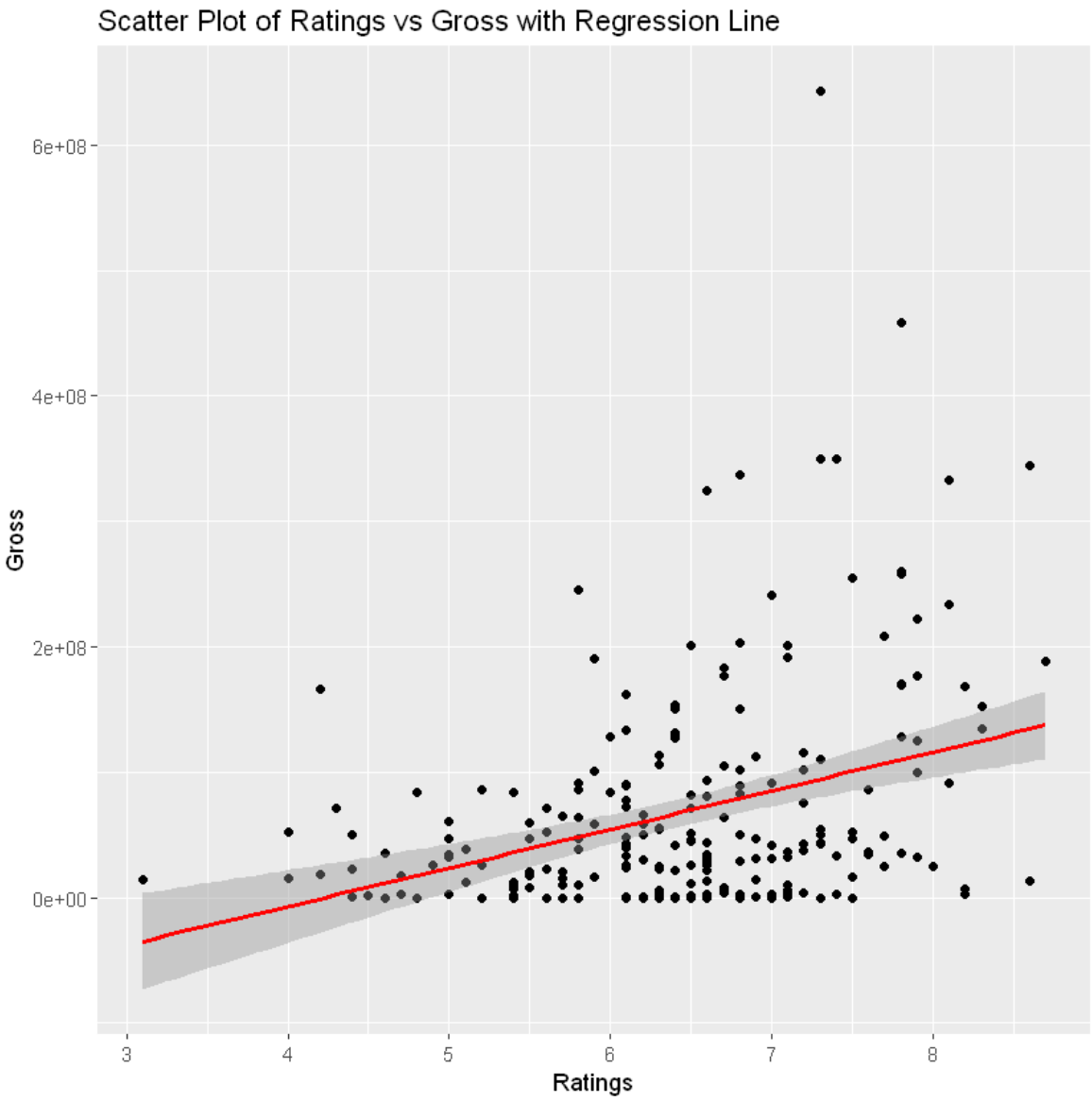
In [53]:
```
ggplot(data, aes(x = Ratings, y = Gross)) +
  geom_bin2d() +
  labs(title = "2D Bin Chart of Ratings vs Gross",
       x = "Ratings",
       y = "Gross") +
  scale_fill_continuous(type = "viridis"
```

```
Error in parse(text = x, srcfile = src): <text>:7:0: unexpected end of input
5:        y = "Gross") +
6:   scale_fill_continuous(type = "viridis"
  ^
Traceback:
```

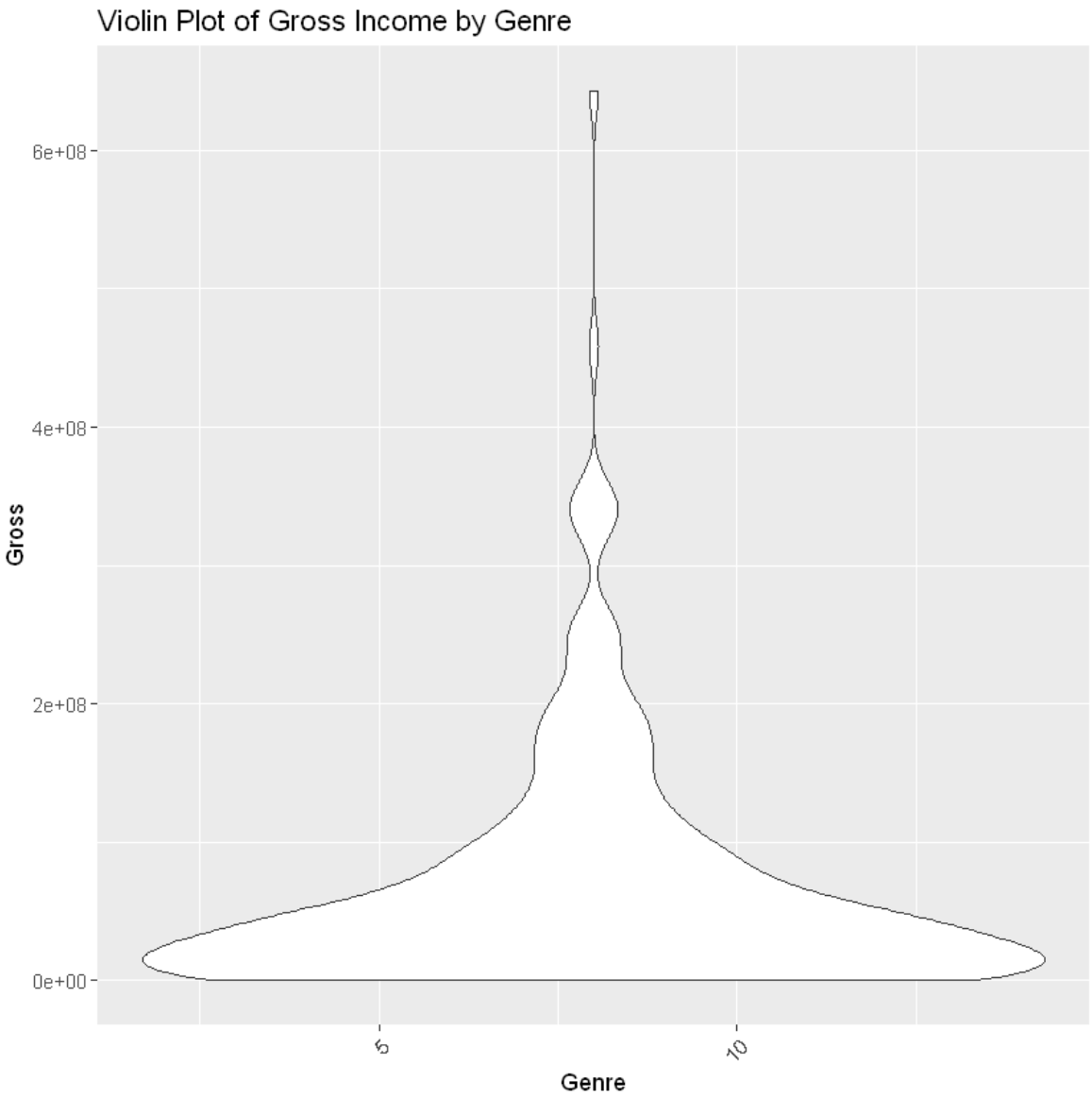
In [54]:
```
ggplot(data, aes(x = Ratings, y = Gross)) +
  geom_hex() +
  labs(title = "Hex Chart of Ratings vs Gross",
       x = "Ratings",
       y = "Gross") +
  scale_fill_continuous(type = "viridis")
```



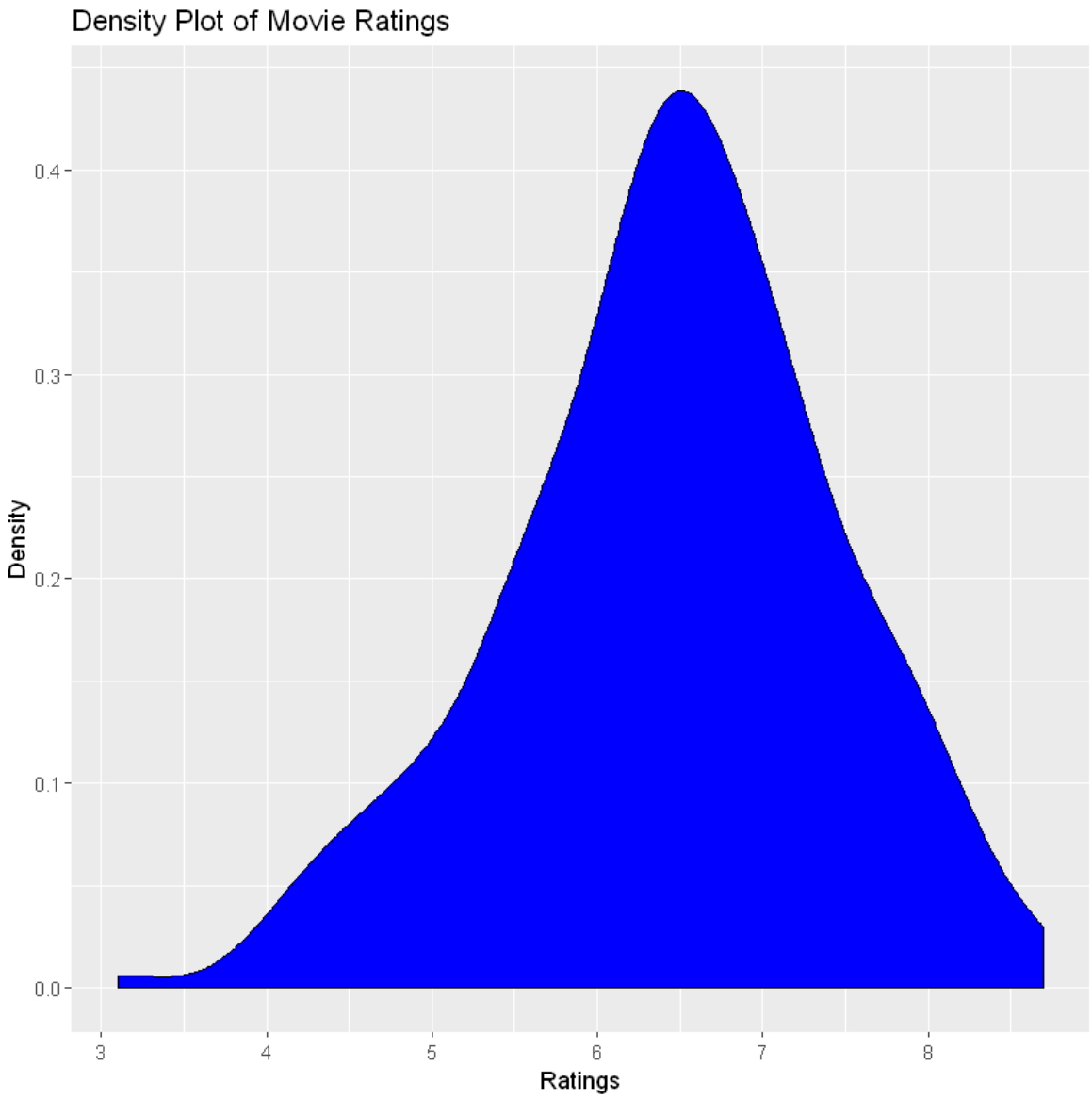
In [55]:
```
ggplot(data, aes(x = Ratings, y = Gross)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red") +
  labs(title = "Scatter Plot of Ratings vs Gross with Regression Line",
       x = "Ratings",
       y = "Gross")
```



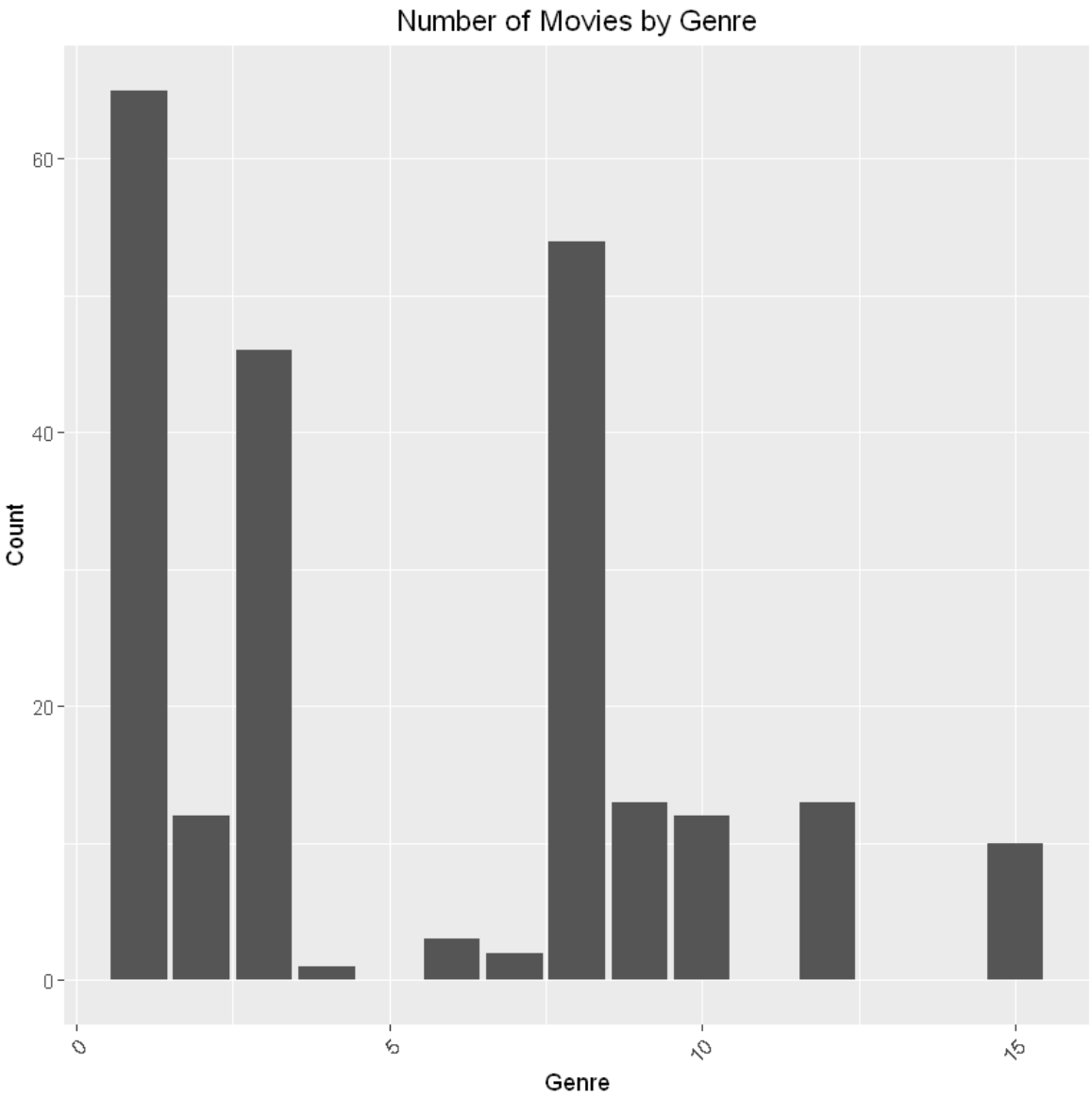
In [56]:
```
ggplot(data, aes(x = Genre, y = Gross)) +
  geom_violin() +
  labs(title = "Violin Plot of Gross Income by Genre",
       x = "Genre",
       y = "Gross") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



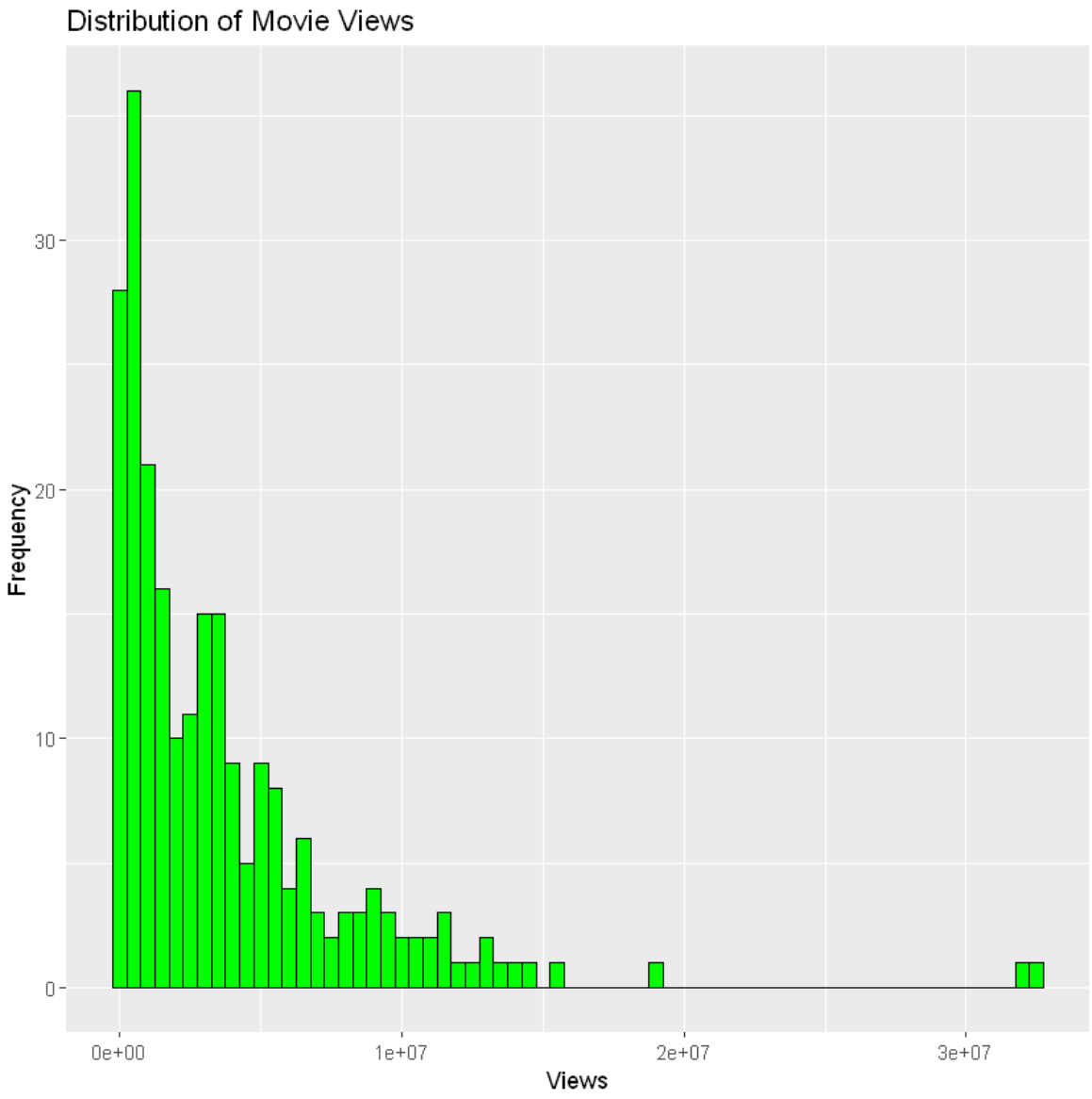
In [57]:
```
ggplot(data, aes(x = Ratings)) +
  geom_density(fill = "blue") +
  labs(title = "Density Plot of Movie Ratings",
       x = "Ratings",
       y = "Density")
```



In [58]:
```
ggplot(data, aes(x = Genre, fill = Genre)) +
  geom_bar() +
  labs(title = "Number of Movies by Genre",
       x = "Genre",
       y = "Count",
       fill = "Genre Type") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```

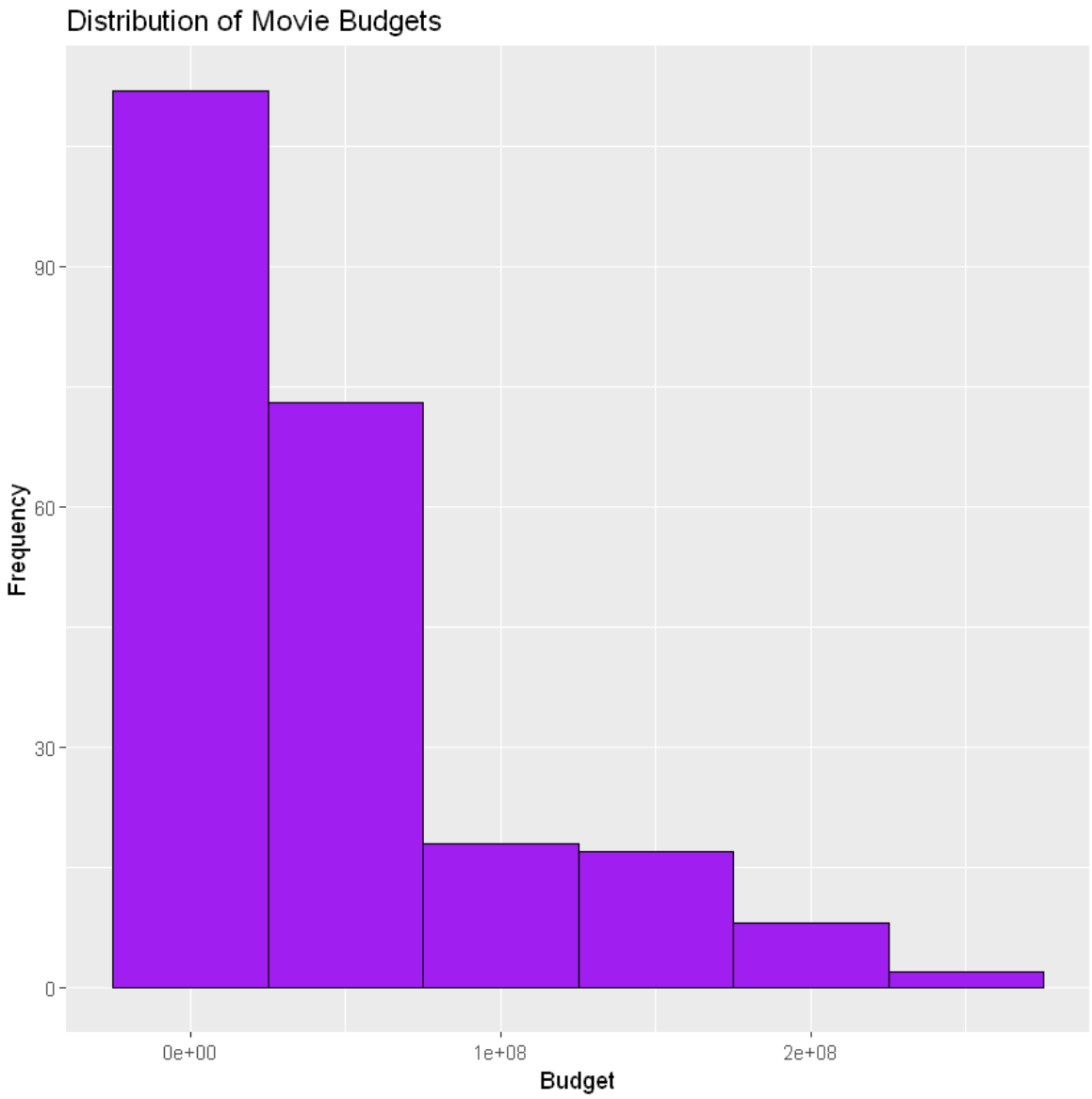


In [59]:
```
ggplot(data, aes(x = Views)) +
  geom_histogram(binwidth = 500000, fill = "green", color = "black") +
  labs(title = "Distribution of Movie Views",
       x = "Views",
       y = "Frequency")
```



In [60]:
```
ggplot(data, aes(x = Budget)) +
  geom_histogram(binwidth = 50000000, fill = "purple", color = "black") +
  labs(title = "Distribution of Movie Budgets",
       x = "Budget",
       y = "Frequency")
```

```
Warning message:
"Removed 1 rows containing non-finite values (stat_bin)."
```


Distribution of Movie Budgets
Frequency
90
60
30
0
0e+00
1e+08
2e+08
Budget

In [ ]: