TicketMaster Analysis of 2021 Concerts

ANLY 511 Group 15 Project Report

Ramdayal, Sangeetha, Chaitanya, Aanchal, Samiksha

Table of contents

1 Introduction

Concerts are a crucial component of an artist's new album promotion. Additionally, it gives fans a chance to see and hear their favorite musicians perform live. Many fans are extremely passionate about the artists they enjoy, and they will go to any lengths to see them perform. Say your favorite performer just made the announcement that they will be making a tour stop in your area. You would probably use your phone or computer to purchase tickets, but what website would you be directed to? Ticketmaster or Live Nation, which is owned by Ticketmaster, are the most likely candidates. Ticketmaster is the largest global ticket marketplace for the majority of concerts and sporting events. Two college students who wanted to improve their theater's ticketing system started it in 1976. The concert industry is worth 51.3 billion dollars in 2022. Every year Ticketmaster sells nearly 500 million tickets for a huge variety of concerts, events, and games, and every year their sites receive more than a billion visits. By far the market leader in the live event and ticketing space, Ticketmaster holds over 70% of the market (and over 80% for live concerts). More commonly, Ticketmaster is recognized for managing artists, distributing ticket-selling software, and selling tickets. Because of the contracts, they have with sporting and concert venues, the majority of artists are compelled to use Ticketmaster in order to sell tickets. While Live Nation jointly owns a sizable portion of venues, strengthening its monopoly. Through this data analysis, we can confirm the elements that affect the minimum cost of the concert ticket. It will be interesting to compare and contrast the variables that affect the minimum price of the concert ticket sold by Ticketmaster given the number of people compelled to use this platform for the most well-known concerts. We'll be closely examining the impact of the venue's size and population, the event's setting, the artists performing, each artist's level of popularity, the performance day (weekend or weekday), and the minimum price of concert tickets.

The following research questions were developed in view of all of these: 1. Does the size and population of the venue affect the price of the concert ticket? 2. Does the location of the event affect the price? 3. Is the price dependent on the artist? 4. Is the pricing impacted for each artist along with their performance date? 5. Is the mean minimum price of Ed Sheeran the same as the mean minimum price of G-Eazy? 6. Is the pricing of the tickets dependent on the popularity of the artist? 7. Is the pricing impacted if the artist performed on a weekend? 8. What factors influence the pricing of the tickets?

This analysis is important because when a global giant dominates a market, consumers feel as though they have no other options therefore, it is crucial for consumers to be aware of the variables that affect the minimum prices of concert tickets sold by Ticketmaster so they can score cheaper tickets.

2 Importing Libraries

```
# CLEARING THE ENVIRONMENT AND LOADING THE LIBRARIES
rm(list=ls())
suppressPackageStartupMessages({library(tidyverse)
library(ggplot2)
library(tidyr)
```

```
library(corrplot)
library(patchwork)
library(car)
library(MASS)
library(dplyr)
library(caret)})
```

3 Import Data

The dataset has been obtained using the official Ticketmaster API and has 8 feature variables and 1 label variable (minprice). The feature variables include: The city in which the event is taking place, name of the Artist, name of the Venue, Weekend or Weekday, Population of the city, Month, Rating of the Artist and Genre of the Music the Artist is known for.

```
# READING THE DATA
df = read_csv('ticketmaster-data.csv')
```

```
New names:
Rows: 1198 Columns: 10
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(5): city, artist, venue, pop, genre dbl (5): ...1, weekend, month, score,
minprice
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
* `` -> `...1`
```

```
df[1:3, ]
```

```
# A tibble: 3 x 10
   ...1 city        artist   venue       weekend pop   month score genre minpr~1
  <dbl> <chr>       <chr>    <chr>         <dbl> <chr> <dbl> <dbl> <chr>   <dbl>
1     1 Detroit     Sam Hunt Ford Field        0 6887~    10    76 Coun~    39.8
2     2 Minneapolis Sam Hunt Target Fie~       1 4000~     7    76 Coun~    24.8
3     3 Denver      Sam Hunt Sports Aut~       1 6494~     8    76 Coun~    29.8
# ... with abbreviated variable name 1: minprice
```

4 Data Cleaning and Pre-Processing

```
# CLEANING THE DATA
df = df[,!(names(df) %in% c('...1'))]
df = df[df$pop != '???',]
```

```
cat(' The number of unique artists:', length(unique(df$artist)), '\n',
    'The number of unique genres:', length(unique(df$genre)), '\n',
    'The number of unique cities:', length(unique(df$city)), '\n',
    'Number of shows on the weekend =', sum(df$weekend), '\n',
    'Number of shows on the weekdays =', nrow(df) - sum(df$weekend))
```

```
 The number of unique artists: 82
 The number of unique genres: 15
 The number of unique cities: 137
 Number of shows on the weekend = 406
 Number of shows on the weekdays = 525
```

```
# CHECKING FOR MISSING VALUES
colSums(is.na(df))
```

```
    city   artist    venue  weekend      pop    month    score    genre
       0        0        0        0        0        0        0        0
minprice
       0
```

```
# REMOVING OUTLIERS
quartiles <- quantile(df$minprice, probs=c(.25, .75), na.rm = FALSE)
IQR <- IQR(df$minprice)
Lower <- quartiles[1] - 1.5*IQR
Upper <- quartiles[2] + 1.5*IQR
df <- subset(df, df$minprice > Lower & df$minprice < Upper)
cat('The number of rows after Removing Outliers from Minimum Price are:', dim(df)[1])
```

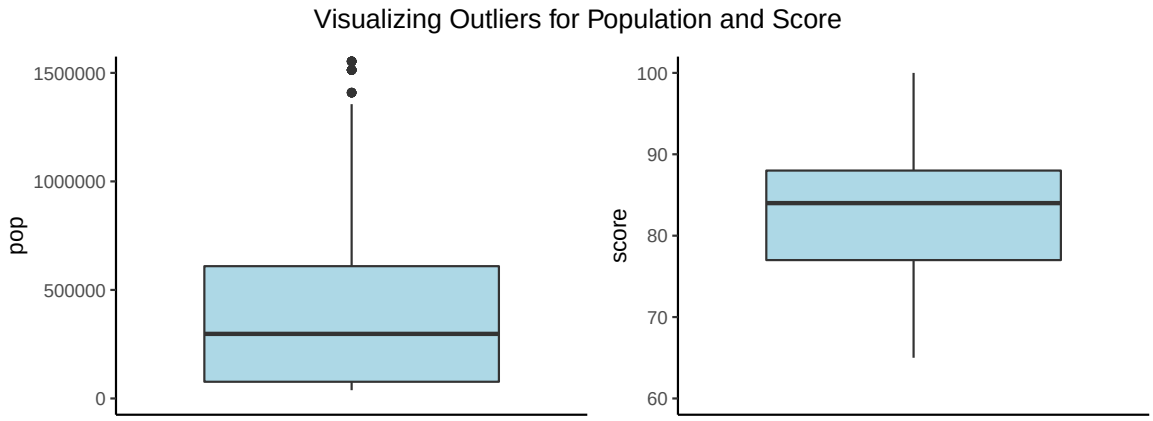
```
The number of rows after Removing Outliers from Minimum Price are: 866
```

```
# CONVERT WEEKEND AND MONTH TO FACTOR and POPULATION TO NUMERIC
df$pop<-as.numeric(df$pop)
df$weekend = as.factor(df$weekend)
df$month = as.factor(df$month)
```

5 Data Visualization

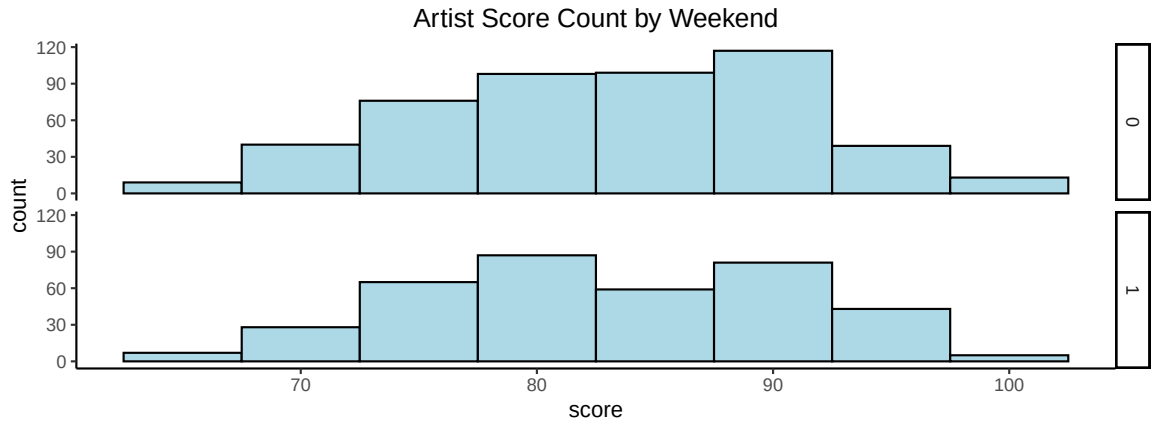
The graphic representation of information and data is known as data visualization. Data visualization tools offer an accessible way to see and understand trends, outliers, and patterns in data by utilizing visual elements like charts, graphs, and maps. Additionally, it offers a great way for staff members or business owners to clearly present data to non-technical audiences. To analyze vast amounts of data and make data-driven decisions, data visualization tools and technologies are crucial in the world of big data.

5.1 *Visualizing Outliers*



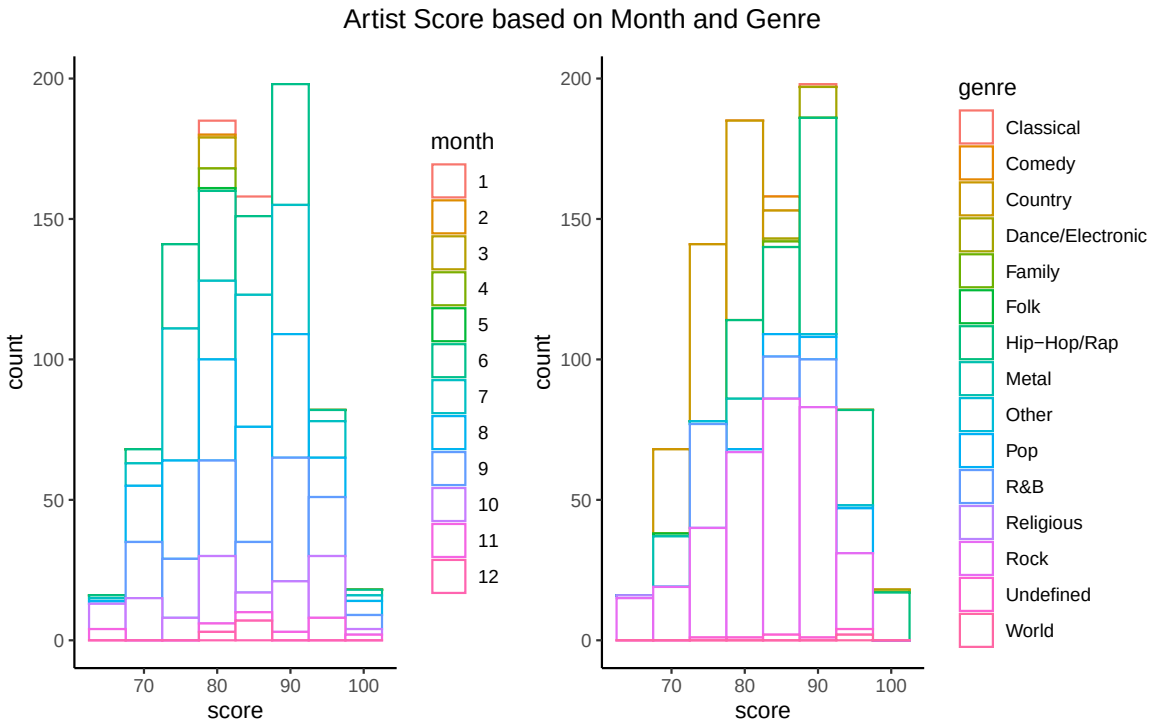
The Mean of Population of the City lies near 350,000 but it has values ranging up to 1,600,000 which leads us to believe that there are outliers in this feature. Since it is a universally verified data and there are cities with Population greater than 1.5 million in the world hence these should not be removed. There are no Outliers in the Rating of the artist and the Mean of the Ratings lie around 84.

5.2 *Distribution of Artist Ratings Weekend-wise*



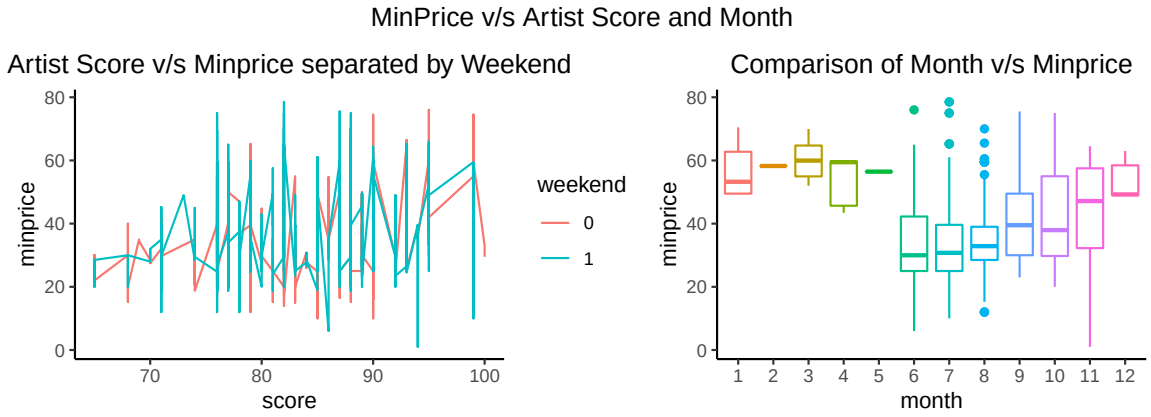
In general, the artists tend to perform more on Weekdays than on Weekends. This might be also because the Weekend comprises of 2 days while Weekdays comprises 5.

5.3 *Distribution of Rating of the Artist based on Month and Genre*



Many artists irrespective of their rating tend to prefer performing from the months May to August (during summer break). Majority of Hip-Hop artists have a higher rating of around 85-95 while Country artists have a lower rating of 65-75. Rock artists are neutral and are equally distributed throughout the rating chart.

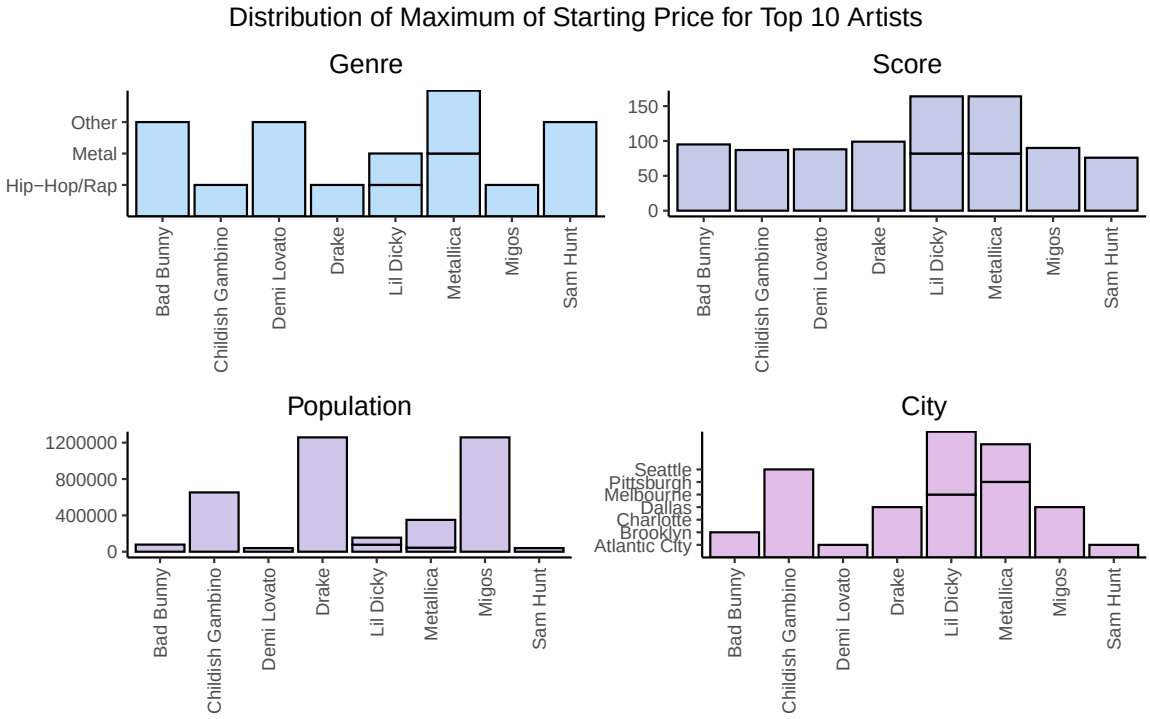
5.4 *Comparison of Starting Price w.r.t. Artist Score and Month*



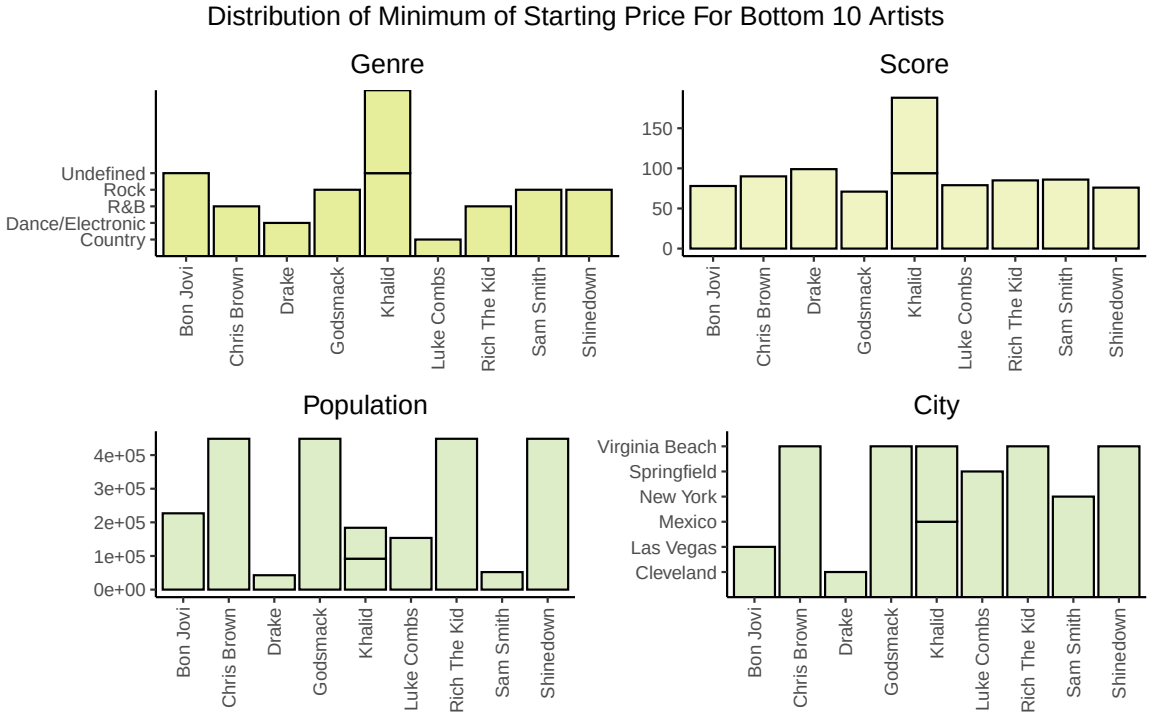
As the Rating of Artists increase, the starting price of an event reaches its peak on Weekdays while for lower artist ratings the starting price reaches its peak on Weekends. Also, the mean Starting price of tickets for months of January to May is higher than other months.

5.5 *Comparison of Artists with Starting Price*

5.5.1 *Comparison of Top 10 Artists with Maximum Starting Price*

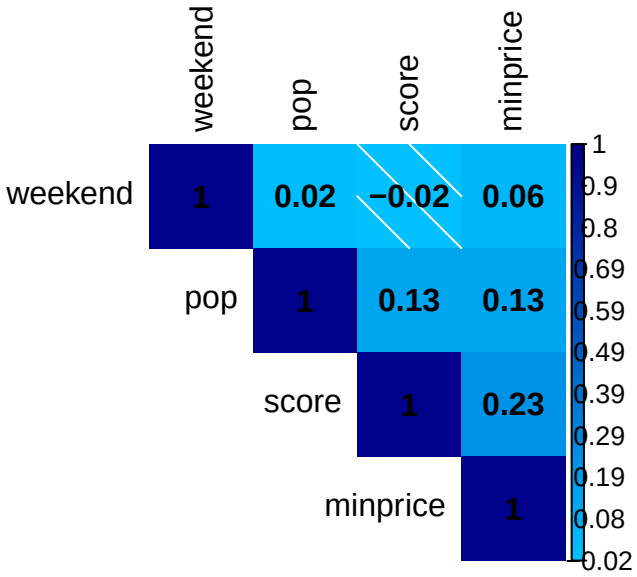


5.5.2 *Comparison of Bottom 10 Artists with Minimum Starting Price*



Genres like Hip-Hop and Metallica have the highest starting price while R&B, Rock and Country have the lowest starting price. Also, Cities like Seattle, Pittsburgh and Melbourne have the highest starting price while Virginia, Springfield and New York have the lowest starting price.

5.6 *Correlation Plot between Numeric Features*



None of the attributes are significantly related.

6 Hypothesis Testing

Using sample data, hypothesis testing is done to determine whether a hypothesis is plausible. The test provides evidence that the hypothesis is plausible in considering the available data. An informed guess is first made on the parameter or distribution. The null hypothesis is also known as H0, because it is the default assumption. The opposite of what is said in the null hypothesis is then specified as an alternative hypothesis (designated Ha). Using sample data, the hypothesis-testing technique determines whether or not H0 may be rejected. The statistical conclusion is that the alternative hypothesis Ha is true if H0 is rejected.

In applications involving hypothesis-testing, the p-value offers a suitable foundation for making conclusions. If the null hypothesis is true, the p-value serves as a measure of how likely the sample results are; the lower the p-value, the less likely the sample results. The null hypothesis can be disregarded if the p-value is less than , otherwise,it cannot be rejected. The p-value is often called the observed level of significance for the test. Regression and correlation analysis both use hypothesis tests to assess whether the correlation coefficient and regression connection are statistically significant.

6.1 T-Test

A statistical test called a t test is employed to compare the means of two groups. It is frequently employed in hypothesis testing to establish whether a process or treatment truly affects the population of interest or whether two groups differ from one another. T-tests are used when the data sets follow a normal distribution and have unknown variances.

A t-test is an inferential statistic used to determine if there is a statistically significant difference between the means of two variables. Calculating a t-test requires three fundamental data values including the difference between the mean values from each data set, the standard deviation of each group, and the number of data values. T-tests can be dependent or independent.

The problem statement is established mathematically by using a sample from each of the two sets in the t-test. It assumes that the two means are equal, which is the null hypothesis.Values are computed and compared to the standard values using the formulas. Accordingly, the assumed null hypothesis is either accepted or rejected. If the null hypothesis can be ruled out, it means that the data readings are significant and almost certainly not random.

6.1.1 When to use a t test

A t test can only be used when comparing the means of two groups. The t test is a parametric test of difference, meaning that it makes the same assumptions about your data as other parametric tests. The t test assumes your data:

1. are independent
2. are (approximately) normally distributed
3. have a similar amount of variance within each group being compared.

6.1.1.1 T-test analysis on Ticket Master Data

The two performers who frequently appear in the same cities at the same time are Ed Sheeran and G-Eazy. Thus, it was decided to compare the minimum prices of these two artists. Here, the case study was whether Ed Sheeran's average minimum price is higher than G- Eazy's.

In a new dataset, we only kept the information about the artists we were going to analyze during this process and excluded all other information.

Our assumed Null Hypothesis : The mean difference of minimum price between these two artists is zero.

Alternate Hypothesis : The mean difference of minimum price between these two artists greater than zero.

Mathematically, Let G= mean minimum price of G-Eazy and E= mean minimum price of Ed Sheeran. Null hypothesis H0 : E− G = 0 Alternative hypothesis Ha : E− G > 0.

```r
# T-TESTING ANALYSIS
df3 <-df %>% filter(artist %in% c("G-Eazy", "Ed Sheeran"))
Ed_Sheeran <- subset(df3, select=minprice,subset=artist=="Ed Sheeran", drop=T)
G_Eazy <- subset(df3, select=minprice, subset=artist=="G-Eazy", drop=T)
t.test(Ed_Sheeran, G_Eazy, alt="greater")
```

```
    Welch Two Sample t-test

data:  Ed_Sheeran and G_Eazy
t = 5.8615, df = 18.387, p-value = 6.879e-06
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 5.40113     Inf
sample estimates:
mean of x mean of y
 37.50000  29.83333
```

The output provides, - An explanation of what is being compared called the "data" in the output table. - A t-value, in this case t = 5.8615 - The degrees of freedom: 18.387. The number of "free" data points in a test that can be used for comparisons is represented by the degrees of freedom, which is connected to your sample size. The more degrees of freedom we have, the more accurate the statistical analysis will be. - The p value: 6.879e-06 (i.e. 6.8 with 15 zeros in front). This expresses the probability that a t value this large would occur by chance. - A statement of the alternative hypothesis (Ha). In this test, the Ha is that true difference in means is greater than 0. - The 95% confidence interval. This is the range of numbers within which the true difference in means will be 95% of the time. This can be changed from 95% to a larger or smaller value. - The mean minimum price of each group.

Since p value is less than 0.05, At 5% significance level, we have enough evidence to reject the null hypothesis. Therefore there is strong evidence that the mean minimum price of Ed Sheeran and G-Eazy were not the same but in fact the average minimum price of Ed Sheerant were greater than the average minimum price of G-Eazy.

Bootstrapping Mean Test:

To estimate the variability in a statistic of interest, bootstrapping involves sampling with replacement from observed data. Bootstrapping is a statistical technique that generates several simulated samples from a single dataset. With this method, standard errors, confidence intervals, and hypothesis testing can all be calculated.

The bootstrap is frequently used to evaluate the precision of an estimate based on a sample of data from a larger population. Consider the sample mean. Drawing numerous different sample means is the most effective technique to learn how they behave.

One method of resampling is to generate a proxy universe based solely on our sample by repeatedly replicating the sample data. Since the sample typically contains all the information we have about the population that gave rise to it, it is sometimes the best place

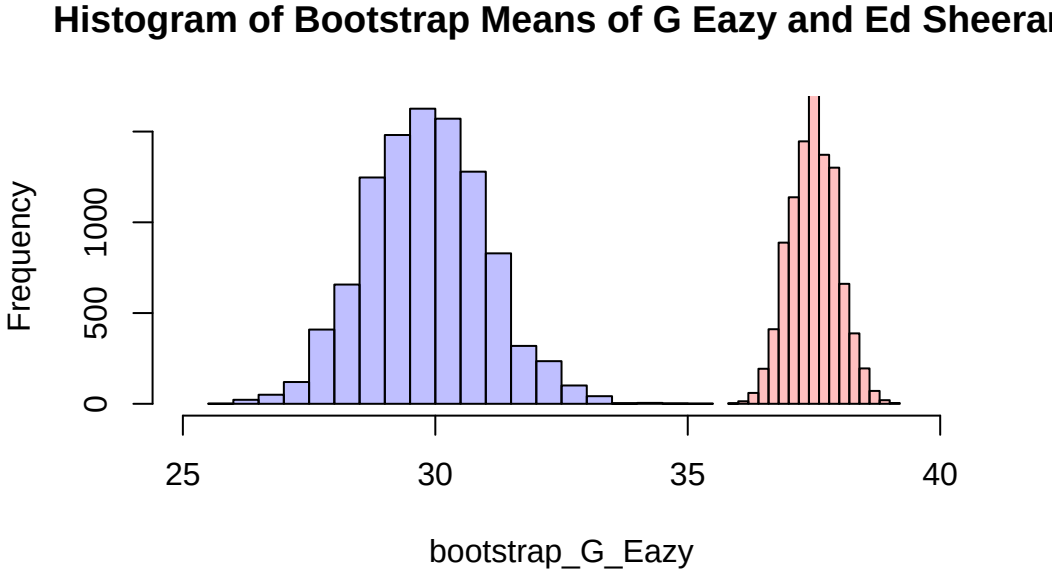
to begin when building an artificial proxy universe from which we can extract resamples and study the distribution of the statistic of interest.

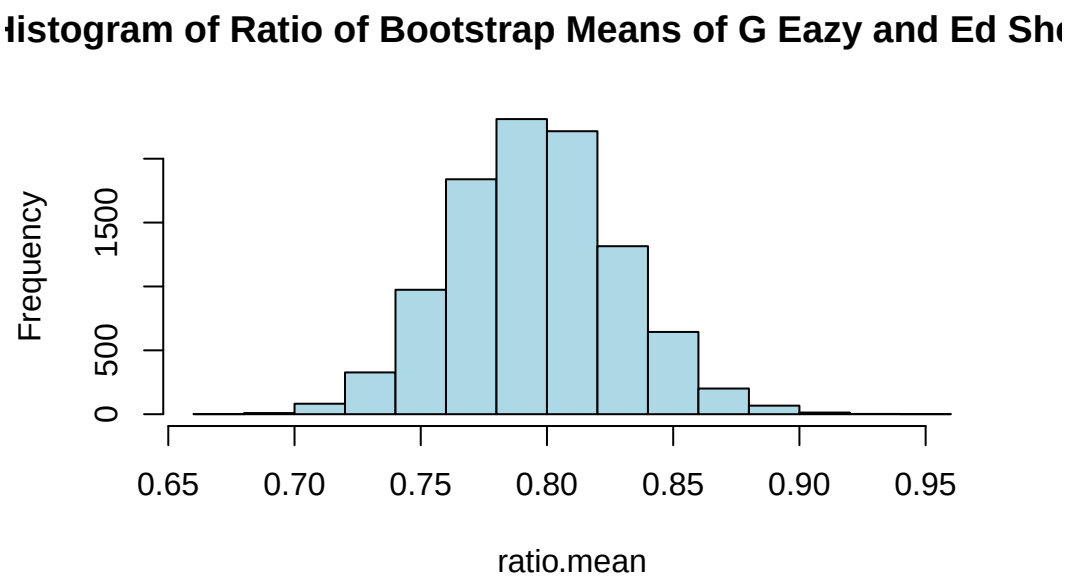
In this case, we replicated our data 10000 times to check the results, the mean ratio was found from bootstrap G-Eazy and Bootstrap Ed Sheeran and we calculated the 95% confidence interval for the difference between the average means of these two artists and obtained the results that at 95% bootstrap percentile interval,mean ratio of G-Eazy's minimum price to Ed Sheeran's minimum price is between 0.7332200 and 0.8629818.

```
#BOOTSTRAP TEST
G_Eazy <- df3[df3$artist == 'G-Eazy',]$minprice
Ed_Sheeran <- df3[df3$artist == 'Ed Sheeran',]$minprice

n_G_Eazy <- length(G_Eazy)
n_Ed_Sheeran <- length(Ed_Sheeran)

bootstrap_G_Eazy <- replicate(10000,mean(sample(G_Eazy , n_G_Eazy , replace = T)))
bootstrap_Ed_Sheeran <- replicate(10000,mean(sample(Ed_Sheeran , n_Ed_Sheeran , replace =

ratio.mean <- bootstrap_G_Eazy/bootstrap_Ed_Sheeran
quantile(ratio.mean, c(0.025, 0.975))
```

```
     2.5%     97.5%
0.7326774 0.8623190
```

At 95% bootstrap percentile interval,mean ratio of G-Eazy's minimum price to Ed Sheeran's minimum price is between 0.7332200 and 0.8629818.





6.2 Chi-Square Test:

Whether there is a statistically significant association between categorical variables is determined by the Chi-square test of independence. Is there a relationship between the values of one category variable and the values of other categorical variables? This is answered by a hypothesis test. Chi-square test has two hypothesis; Null hypothesis : The categorical variables do not have any relationships with one another. Knowing the value of one variable does not make it easier to predict the value of another. Alternate hypothesis : The category variables are related to one another. It does make it easier to predict the value of another variable if you are aware of the value of one.

A p-value for a Chi-square test below or equal to the significance level means there is enough data to draw the conclusion that the observed distribution differs from the expected distribution. It is clear that there is a connection between the categorical variables.

6.2.0.1 Chi-square test analysis on Ticket Master Data

We considered two categorical variables 'Artist' and 'City' to find the correlation between them. We stated the hypothesis; Null Hypothesis H0 : The variables artist and city are independent to each other Alternate Hypothesis Ha : The variables artist and city are not independent to each other.

```
# CHI-SQUARE ANLYSIS
chisq.table = table(df[, c('artist', 'city')])
chisq.test(chisq.table)
```

```
Warning in chisq.test(chisq.table): Chi-squared approximation may be incorrect
```

```

    Pearson's Chi-squared test
```

```
data:  chisq.table
X-squared = 12222, df = 10395, p-value < 2.2e-16
```

From the output we see that the p-value is less than the significance level of 5%. If the p-value is less than the significance level, we can reject the null hypothesis. In our context, rejecting the null hypothesis for the Chi-square test of independence means that there is a significant relationship between the Artist and the city.

Yates Continuity Correction:

This Chi square test assumes that the continuous Chi-Square distribution may adequately approximate the discrete probability of the frequencies in a contingency table. The test statistic that results is typically skewed upwards because this assumption is prone to being a little inaccurate.

To correct for this bias we can apply Yate's continuity correction, which applies the following correction to the X2 formula:

$X2 = \Sigma(|Oi\text{-}Ei| - 0.5)2 \ / \ Ei$

where O is the observed value, E is the expected value and Σ is the sum value.

```r
# YATES CORRECTION
chi_square_test1 = chisq.test(chisq.table, correct = TRUE, simulate.p.value = TRUE);
chi_square_test1
```

```
    Pearson's Chi-squared test with simulated p-value (based on 2000
    replicates)

data:  chisq.table
X-squared = 12222, df = NA, p-value = 0.002499
```

According to the test, the p value that corresponds to the test is p = 0.002999, which is less than 5% significant value, so we can reject the null hypothesis.

6.3 ANOVA Test

6.3.1 One Way ANOVA:

When testing an hypothesis with a categorical explanatory variable and a quantitative response variable, the tool normally used in statistics is Analysis of Variances, also called ANOVA. We are performing an ANOVA test using the R programming language, to a dataset of ticketmaster minimum price across artists. The objective of the ANOVA test is to analyse if there is a (statistically) significant difference in minimum price of the concert tickets, between different artists. In other words, I am interested to see whether minimum price of concert tickets are more likely to change with the artists' performing. The Hypothesis is: Null Hypothesis: There is no effect of the Artist on the Minimum Price. Alternative Hypothesis: There is an effect of the Artist on the Minimum Price. Here the artists are the explanatory variable and minimum price is the response variable.

```r
# ONE WAY ANOVA TESTING
anova_minprice_pop = aov(minprice ~ artist, data = df)
summary(anova_minprice_pop)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
artist       77  93849  1218.8   17.48 <2e-16 ***
Residuals   788  54937    69.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that, our F value is 17.48, and p-value is very low too. In other words, the variation of minimum price of tickets means among different artists is much larger than the variation of minimum price of tickets within each artist performing, and our p-value is less than 0.05, Hence we reject the null hypothesis H0 and can conclude that there is a significant relationship between artists and minimum price of concert tickets.

6.3.2 Two Way ANOVA:

The two-way ANOVA test is used to simultaneously compare the effects of two grouping variables on a response variable at the same time. The objective of the ANOVA test is to analyse if there is a (statistically) significant difference in minimum price of the concert tickets, between different cities and different days (i.e. weekend or weekday). In other words, we are interested to see whether minimum price of concert tickets are more likely to change with the cities and weekends. The Hypothesis is: Null Hypothesis: There is no effect of the City and Weekend on the Minimum Price. Alternative Hypothesis: There is an effect of the City and Weekend on the Minimum Price. Here the city and weekend are the explanatory variable and minimum price is the response variable.

```r
# TWO WAY ANOVA
df$city<- as.factor(df$city)
df$weekend<-as.factor(df$weekend)
anova1<-aov(minprice~weekend+city, data=df)
summary(anova1)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
weekend       1    497   497.4   3.920 0.0481 *
city        135  55785   413.2   3.256 <2e-16 ***
Residuals   729  92504   126.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova2<- aov(minprice~weekend*city, data=df)
summary(anova2)
```

```
              Df Sum Sq Mean Sq F value  Pr(>F)
weekend        1    497   497.4   4.150 0.04205 *
city         135  55785   413.2   3.447 < 2e-16 ***
weekend:city  80  14707   183.8   1.534 0.00309 **
Residuals    649  77797   119.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that, my F value is 3.92 for weekend and 3.25 for city, and p-value is very low too. In other words, the variation of minimum price of tickets means among different cities and different days is much larger than the variation of minimum price of tickets within each city and weekend or weekday performances. The p-value is less than 0.05, Hence we reject the null hypothesis H0 and can conclude that there is a significant relationship between cities and weekends to the minimum price of concert tickets. We go ahead to check if you think these two variables will interact to create a synergistic effect. Here, we can see that city, weekend and their interaction effect are all significant since their p-values are less than 0.05 and can conclude that there is a significant relationship between cities, weekends and their interaction effect to the minimum price of concert tickets.

7 Linear Regression

- A fundamental and widely used form of predictive analysis is linear regression. Regression analysis' main goal is to look at two things:

1. Is it possible to accurately predict an outcome (dependent) variable using a set of predictor variables?
2. Which particular variables—as shown by the size and sign of the beta estimates—are highly significant predictors of the outcome variable, and how do they affect the outcome variable?

- The relationship between one dependent variable and one or more independent variables is explained using these regression estimates. The following formula represents the regression equation's most basic version with one dependent variable and one independent variable:

$$\mathrm{Y} = \alpha + \beta_1(\mathrm{X}_1) + \beta_2(\mathrm{X}_2) + ... + \beta_n(\mathrm{X}_\mathrm{n})$$

where Y = Estimated Dependent Variable Score, Alpha = Intercept, Beta = Regression Coefficient, and X = Score of the Independent Variable.

- Three major uses for regression analysis are:

1. Determining the strength of predictors
2. Forecasting an effect
3. Trend forecasting

- Types of Linear Regression:

1. Simple linear regression: 1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)
2. Multiple linear regression: 1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)
3. Logistic regression: 1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)
4. Ordinal regression: 1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)
5. Multinomial regression: 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)
6. Discriminant analysis: 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

7.1 Data Pre-Processing

The data needs to pre-processed before fitting it into any linear model.

Converting all the variables to numeric data type:

```
df$artist = as.numeric(as.factor(df$artist))
df$city = as.numeric(as.factor(df$city))
df$venue = as.numeric(as.factor(df$venue))
df$weekend = as.numeric(as.factor(df$weekend))
df$month = as.numeric(as.factor(df$month))
df$genre = as.numeric(as.factor(df$genre))
```

Data Normalization: After converting the data to numeric, all the values have different ranges and values and must be normalized to a similar range. Z-score standardization has been performed on our data:

```
label_df = df[9]
normalized_df <- as.data.frame(scale(df[1:8]))
```

```
final_df = cbind(normalized_df, label_df)
head(final_df)
```

```
        city     artist      venue    weekend        pop      month      score
1 -0.7352715  1.2474916 -0.5936548 -0.8734217  0.2965677  1.2456215 -0.9557091
2  0.3739025  1.2474916  1.0069133  1.1436002 -0.1332865 -0.4504606 -0.9557091
3 -0.7616804  1.2474916  0.8520196  1.1436002  0.2381787  0.1149001 -0.9557091
4 -1.6331743  0.8326369  1.5103178 -0.8734217 -0.6363257  0.6802608 -0.6900554
5 -0.4183646  1.2474916  1.6781193 -0.8734217 -0.5429192 -1.0158213 -0.9557091
6  1.0077163 -0.5963070 -0.8905344 -0.8734217 -0.0861119  0.6802608  0.6382131
       genre minprice
1 -1.5062250    39.75
2 -1.5062250    24.75
3 -1.5062250    29.75
```

```
4  1.0157991    47.00
5 -1.5062250    20.00
6 -0.4974154    29.50
```

Splitting the data into train and test:

```
set.seed(1973)

training.samples <- final_df$minprice %>%
  createDataPartition(p = 0.85, list = FALSE)
train.data  <- final_df[training.samples, ]
test.data <- final_df[-training.samples, ]
cat('The number of rows in training data is:', dim(train.data)[1])
```

```
The number of rows in training data is: 738
```

```
cat('The number of rows in test data is:', dim(test.data)[1])
```

```
The number of rows in test data is: 128
```

7.2 Model 1

Firstly we fit all our independent variables to check which features are infact the most significant i.e., have the most effect on our dependent variable (minprice). The formula for this fit can be given as:

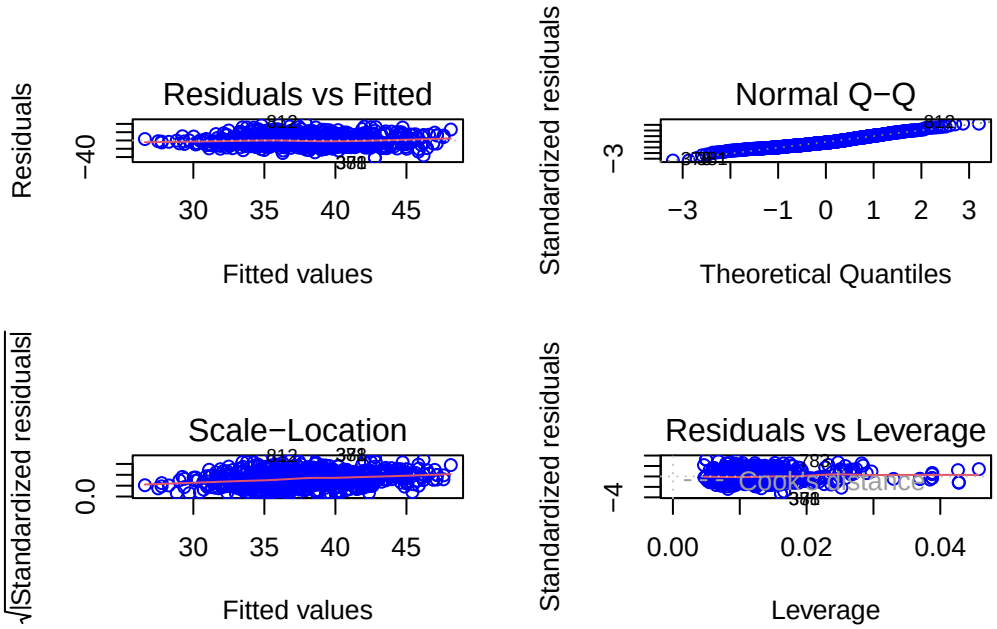
$$\begin{aligned}\text{minprice} = \alpha + \beta_1(\text{city}) + \beta_2(\text{artist}) + \beta_3(\text{venue}) + \\ \beta_4(\text{weekend}) + \beta_5(\text{pop}) + \beta_6(\text{month}) + \beta_7(\text{score}) + \\ \beta_8(\text{genre}) + \epsilon\end{aligned}$$

```
# LINEAR REGRESSION MODEL 1
fit1<-lm(minprice~., data = train.data)
summary(fit1)
```

```
Call:
lm(formula = minprice ~ ., data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-41.779  -8.961  -2.018   8.468  40.623

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  37.7773     0.4670  80.888  < 2e-16 ***
city         -0.3184     0.4695  -0.678  0.49787
artist       -1.3487     0.4758  -2.835  0.00472 **
venue        -1.1292     0.4741  -2.382  0.01748 *
weekend       0.7419     0.4678   1.586  0.11319
pop           1.3130     0.4632   2.834  0.00472 **
month         0.1164     0.4710   0.247  0.80484
score         2.8339     0.4748   5.969 3.74e-09 ***
genre        -0.5117     0.4802  -1.066  0.28692
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.68 on 729 degrees of freedom
Multiple R-squared:  0.08237,   Adjusted R-squared:  0.0723
F-statistic:  8.18 on 8 and 729 DF,  p-value: 1.242e-10
```

```r
par(mfrow=c(2,2))
plot(fit1, col = "blue")
```



Seeing the summary statistics of our first fit we can conclude that *the Artist*, *Venue of the Event*, *Population of the city the event is in* and *the Rating of the Artist* are the main features that help predicting the Starting price of and event.

7.3 Model 2

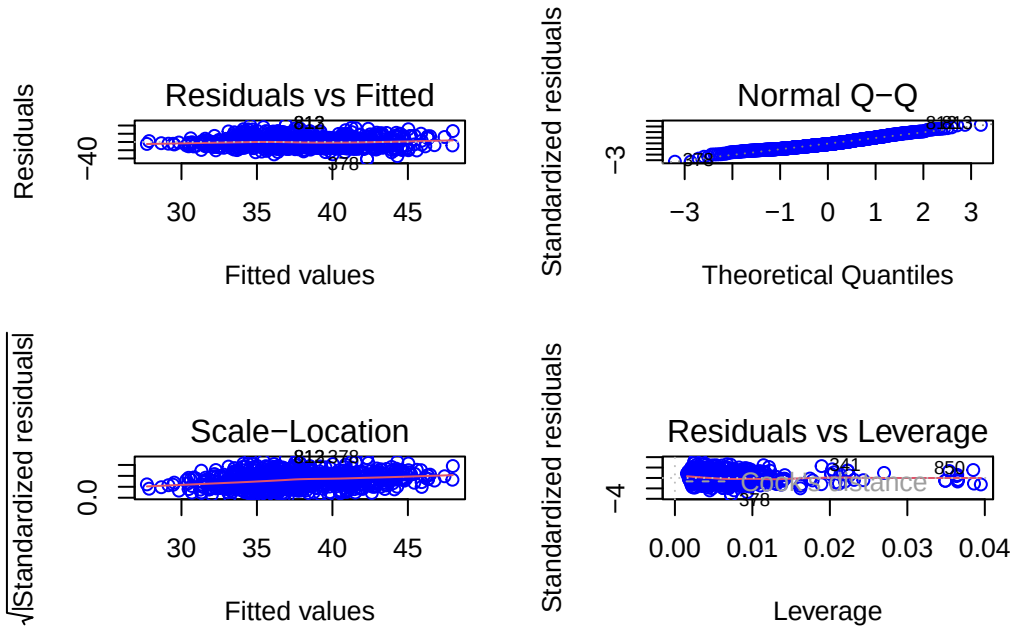
From the previous model we fit the model again with our top 4 features: *the Artist*, *Venue of the Event*, *Population of the city the event is in* and *the Rating of the Artist*. The formula for this fit can be given as:

$$\text{minprice} = \alpha + \beta_1(\text{artist}) + \beta_2(\text{venue}) + \beta_3(\text{pop}) \ + \beta_4(\text{score}) + \epsilon$$

```r
# LINEAR REGRESSION MODEL 2
fit2<-lm(minprice~artist+venue+pop+score, data = train.data)
summary(fit2)
```

```
Call:
lm(formula = minprice ~ artist + venue + pop + score, data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-41.295  -9.101  -1.809   8.054  41.609

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.7581     0.4668  80.880  < 2e-16 ***
artist       -1.4026     0.4713  -2.976  0.00302 **
venue        -1.1211     0.4734  -2.368  0.01814 *
pop           1.2892     0.4612   2.796  0.00532 **
score         2.7379     0.4697   5.828 8.38e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.68 on 733 degrees of freedom
Multiple R-squared:  0.07719,   Adjusted R-squared:  0.07215
F-statistic: 15.33 on 4 and 733 DF,  p-value: 4.798e-12
```

```r
par(mfrow=c(2,2))
plot(fit2, col = "blue")
```



There is not a lot of difference between the first and second fit models in Adjusted R-Squared values. Hence more complex models need to be applied on this data to get better results.

7.4 Model 3

A quadratic regression is the process of finding the equation of the parabola that best fits a set of data. The formula for this fit can be given as:

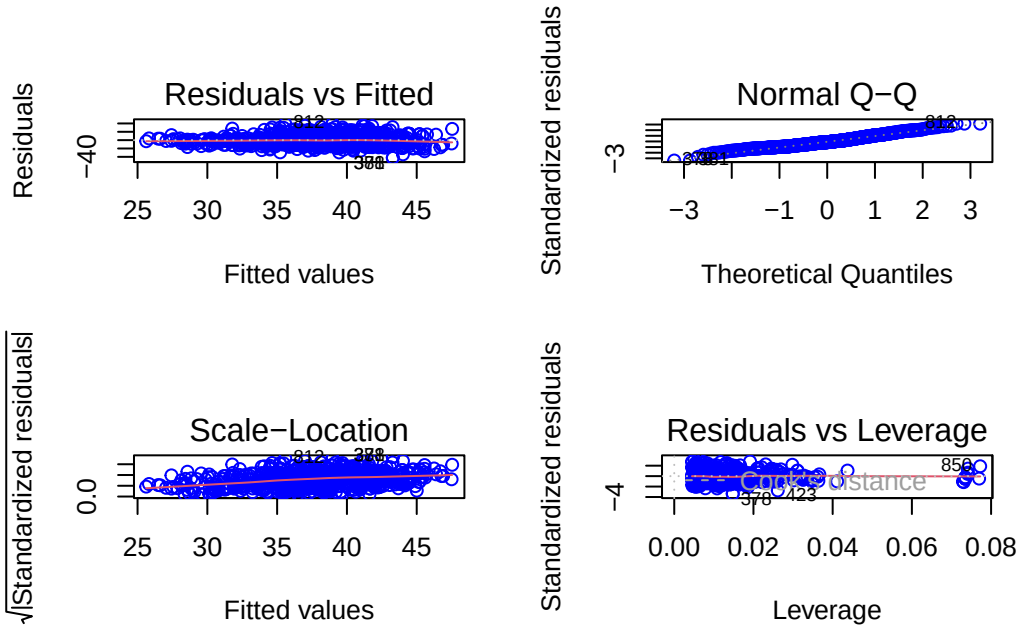
$$\begin{aligned}\text{minprice} = \alpha + \beta_1(\text{artist}) + \beta_2(\text{artist}^2) + \beta_3(\text{venue}) + \\ \beta_4(\text{venue}^2) + \beta_5(\text{pop}) + \beta_6(\text{pop}^2) + \beta_7(\text{score}) + \beta_8(\text{score}^2) + \epsilon\end{aligned}$$

```
# LINEAR REGRESSION MODEL 3 - Quadratic
lm.fitquad = lm(minprice~artist+I(artist^2)+venue+I(venue^2)+
                  pop+I(pop^2)+score+I(score^2), data = train.data)
summary(lm.fitquad)
```

```
Call:
lm(formula = minprice ~ artist + I(artist^2) + venue + I(venue^2) +
    pop + I(pop^2) + score + I(score^2), data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-42.289  -9.284  -1.614   7.893  39.561

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  40.6488     0.9762  41.640  < 2e-16 ***
artist       -1.5446     0.4794  -3.222  0.00133 **
I(artist^2)  -1.8632     0.5303  -3.513  0.00047 ***
venue        -1.0470     0.4844  -2.161  0.03099 *
I(venue^2)   -0.4811     0.5212  -0.923  0.35626
pop           1.9592     0.9218   2.125  0.03389 *
I(pop^2)     -0.2426     0.2628  -0.923  0.35628
score         2.9529     0.4852   6.086 1.87e-09 ***
I(score^2)   -0.3532     0.3937  -0.897  0.36989
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.58 on 729 degrees of freedom
Multiple R-squared:  0.09613,   Adjusted R-squared:  0.08621
F-statistic: 9.692 on 8 and 729 DF,  p-value: 7.891e-13
```

```r
par(mfrow=c(2,2))
plot(lm.fitquad, col = "blue")
```



The Quadratic Regression model increased the Adjusted R-Squared slightly from the previous two models.

7.5 Model 4

As a special case of multiple linear regression, polynomial regression is a type of linear regression that estimates the relationship as an nth degree polynomial. The formula for this fit can be given as:

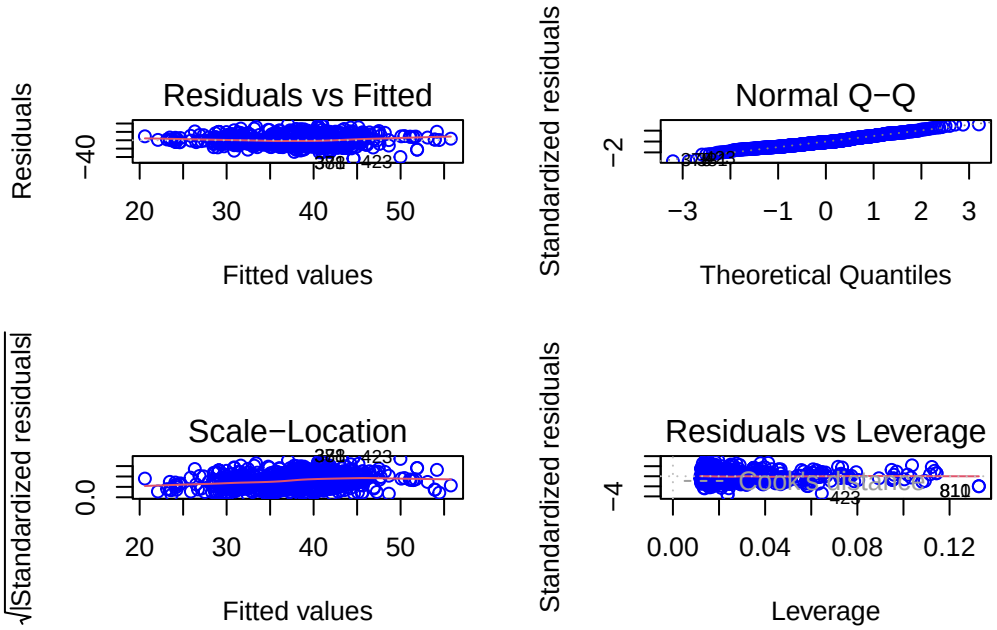
$$\begin{aligned}\text{minprice} = \alpha &+ \beta_1(\text{artist}) + \beta_2(\text{artist}^2) + \beta_3(\text{artist}^3) + \\ &\beta_4(\text{artist}^4) + \beta_5(\text{artist}^5) + \beta_6(\text{venue}) + \beta_7(\text{venue}^2) + \\ &\beta_8(\text{venue}^3) + \beta_9(\text{venue}^4) + \beta_{10}(\text{venue}^5) + \beta_{11}(\text{pop}) + \\ &\beta_{12}(\text{pop}^2) + \beta_{13}(\text{pop}^3) + \beta_{14}(\text{pop}^4) + \beta_{15}(\text{pop}^5) + \\ &\beta_{16}(\text{score}) + \beta_{17}(\text{score}^2) + \beta_{18}(\text{score}^3) + \beta_{19}(\text{score}^4) + \\ &\beta_{20}(\text{score}^5) + \epsilon\end{aligned}$$

```r
# LINEAR REGRESSION MODEL 4 - Polynomial
lm.fitpoly = lm(minprice~poly(artist, 5)+poly(venue, 5)+
                  poly(pop, 5)+poly(score, 5), data = train.data)
summary(lm.fitpoly)
```

```
Call:
lm(formula = minprice ~ poly(artist, 5) + poly(venue, 5) + poly(pop,
    5) + poly(score, 5), data = train.data)

Residuals:
    Min      1Q  Median      3Q     Max
-43.585  -8.357  -1.389   8.270  37.933

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        37.848      0.447  84.668  < 2e-16 ***
poly(artist, 5)1  -35.601     12.934  -2.753 0.006063 **
poly(artist, 5)2  -36.077     12.895  -2.798 0.005283 **
poly(artist, 5)3  -46.366     12.859  -3.606 0.000333 ***
poly(artist, 5)4   22.146     12.586   1.760 0.078901 .
poly(artist, 5)5   44.975     12.451   3.612 0.000325 ***
poly(venue, 5)1   -28.976     12.977  -2.233 0.025868 *
poly(venue, 5)2    -7.170     12.486  -0.574 0.565986
poly(venue, 5)3   -25.604     12.476  -2.052 0.040514 *
poly(venue, 5)4   -13.997     12.475  -1.122 0.262245
poly(venue, 5)5   -19.268     12.863  -1.498 0.134573
poly(pop, 5)1      26.852     12.871   2.086 0.037307 *
poly(pop, 5)2     -11.024     12.796  -0.862 0.389237
poly(pop, 5)3       9.505     12.552   0.757 0.449109
poly(pop, 5)4     -25.122     12.601  -1.994 0.046559 *
poly(pop, 5)5     -21.898     12.489  -1.753 0.079952 .
poly(score, 5)1    90.261     12.929   6.981 6.67e-12 ***
poly(score, 5)2   -10.219     12.701  -0.805 0.421322
poly(score, 5)3    59.374     12.693   4.678 3.47e-06 ***
poly(score, 5)4    11.132     12.885   0.864 0.387886
poly(score, 5)5    -1.786     12.847  -0.139 0.889452
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 12.14 on 717 degrees of freedom
Multiple R-squared:  0.1718,    Adjusted R-squared:  0.1486
F-statistic: 7.434 on 20 and 717 DF,  p-value: < 2.2e-16
```

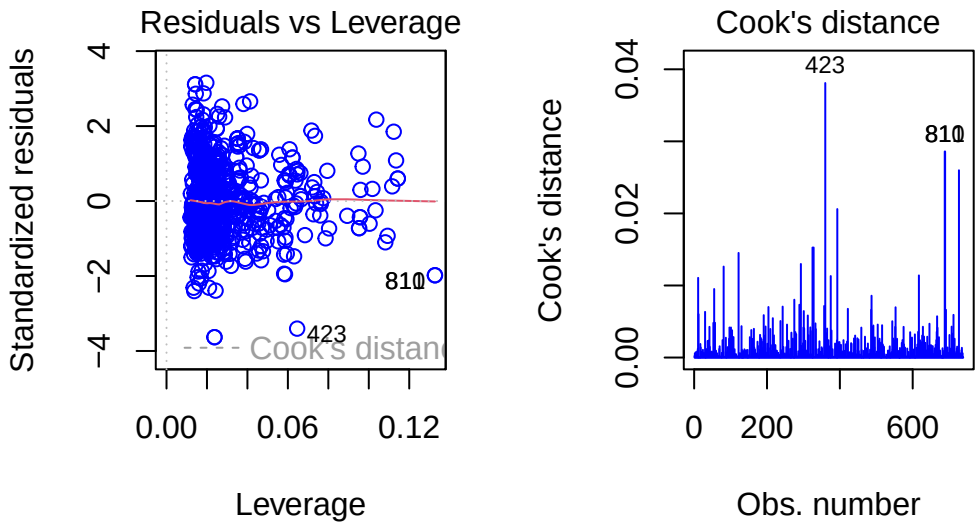
```
par(mfrow=c(2,2))
plot(lm.fitpoly, col='blue')
```



From the graphs we can see that there are number of outliers present in our data but even so our Adjusted R-Squared value has increased significantly from the previous 3 models.

7.5.1 Checking for Outliers

```
par(mfrow=c(1,2))
plot(lm.fitpoly, which= 5, col = "blue")
plot (lm.fitpoly, which = 4, col = "blue")
```



The Polynomial model has some outliers. After checking these values from our original dataframe we can confirm that these are infact outliers and must be removed. One of the examples of the outliers is Outlier number 423 which has minimum price has 1$ as the starting price of the event even though the Artist of the event is top tier and has a very high rating.

7.6 Model 5

Polynomial Regression is sensitive to outliers so the presence of one or two outliers can also badly affect the performance. Hence we remove these outliers using Cook's Distance. A general rule of thumb is to investigate any point that is more than 4x the mean of all the distances. The formula for this fit is the same our previous model since the model parameter is not changing only outliers are being removed.

```
cooksD <- cooks.distance(lm.fitpoly)
influential <- cooksD[(cooksD > (4 * mean(cooksD, na.rm = TRUE)))]
names_of_influential <- names(influential)
outliers <- train.data[names_of_influential,]
train.data_without_outliers <- train.data %>% anti_join(outliers)
```

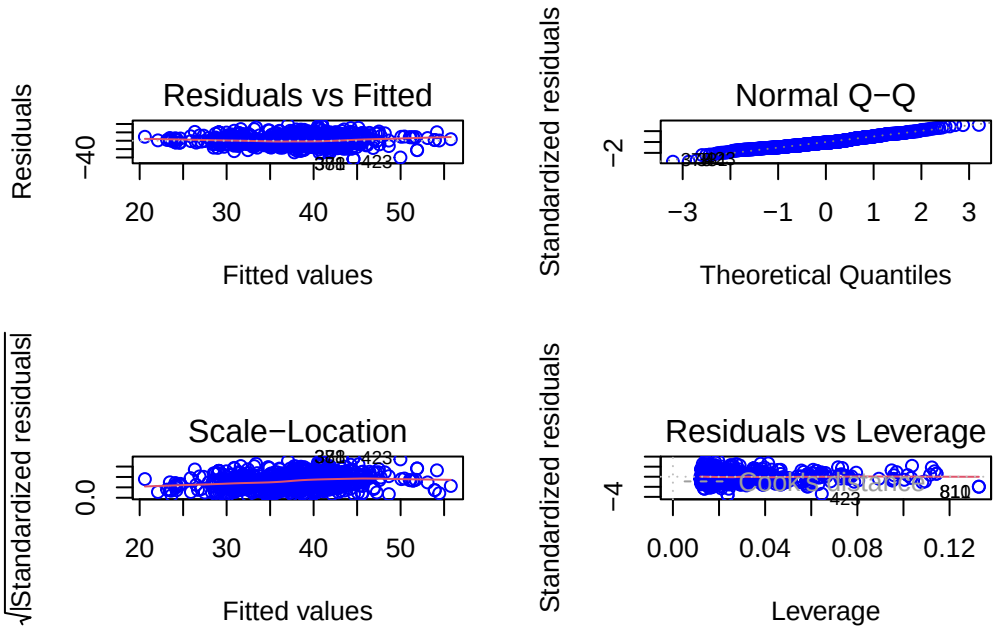
```
Joining, by = c("city", "artist", "venue", "weekend", "pop", "month", "score",
"genre", "minprice")
```

```
lm.fitpoly1 = lm(minprice~poly(artist, 5)+poly(venue, 5)+
                   poly(pop, 5)+poly(score, 5), data = train.data_without_outliers)
summary(lm.fitpoly1)
```

```
Call:
lm(formula = minprice ~ poly(artist, 5) + poly(venue, 5) + poly(pop,
```

```
    5) + poly(score, 5), data = train.data_without_outliers)

Residuals:
    Min      1Q  Median      3Q     Max
-29.433  -7.528  -1.089   7.179  31.327

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       37.4216     0.4072  91.901  < 2e-16 ***
poly(artist, 5)1 -33.5096    11.6592  -2.874 0.004178 **
poly(artist, 5)2 -38.8588    11.5571  -3.362 0.000816 ***
poly(artist, 5)3 -44.9909    11.5287  -3.903 0.000105 ***
poly(artist, 5)4  30.4104    11.1708   2.722 0.006647 **
poly(artist, 5)5  44.3395    11.1141   3.989 7.33e-05 ***
poly(venue, 5)1  -28.0956    11.6036  -2.421 0.015724 *
poly(venue, 5)2  -14.4232    11.1056  -1.299 0.194474
poly(venue, 5)3  -21.4694    11.1327  -1.929 0.054205 .
poly(venue, 5)4  -12.5566    11.0785  -1.133 0.257435
poly(venue, 5)5  -26.5520    11.5342  -2.302 0.021633 *
poly(pop, 5)1     12.5417    11.5114   1.090 0.276316
poly(pop, 5)2    -10.7507    11.3843  -0.944 0.345326
poly(pop, 5)3      7.5211    11.2406   0.669 0.503659
poly(pop, 5)4    -30.1257    11.2887  -2.669 0.007795 **
poly(pop, 5)5    -23.8628    11.2145  -2.128 0.033705 *
poly(score, 5)1   95.4484    11.5830   8.240 8.73e-16 ***
poly(score, 5)2   -2.2195    11.3911  -0.195 0.845573
poly(score, 5)3   63.5874    11.4138   5.571 3.64e-08 ***
poly(score, 5)4   23.8872    11.3985   2.096 0.036480 *
poly(score, 5)5   14.0832    11.4698   1.228 0.219923
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.82 on 685 degrees of freedom
Multiple R-squared:  0.2268,    Adjusted R-squared:  0.2042
F-statistic: 10.04 on 20 and 685 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(lm.fitpoly, col='blue')
```



Removing outliers has increased the model's Adjusted R-Squared even though the model's parameters have not been changed.

7.7 Summary of Evaluation Metrics of the 5 Models:

	RMSE	R2	F stat	Adj R^2	RSE	Models
1	11.81528	0.1663346	8.179749	0.07230016	12.67653	Fit1
2	11.92343	0.1531811	15.327329	0.07214987	12.67756	Fit2
3	12.03232	0.1335647	9.691527	0.08621133	12.58113	Quadratic
4	11.88384	0.1570334	7.434070	0.14864753	12.14371	Polynomial
5	12.20556	0.1249189	10.044927	0.20419781	10.81945	Polynomial Without Outliers

After comparing all the models, the Polynomial Model without Outliers has the highest Adjusted R-Squared, hence we choose that model as our best model.

8 Conclusion

1. Many artists irrespective of their rating tend to prefer performing from the months May to August (during summer break).
2. Majority of Hip-Hop artists have a higher rating of around 85-95 while Country artists have a lower rating of 65-75. Rock artists are neutral and are equally distributed throughout the rating chart.
3. Genres like Hip-Hop and Metallica have the highest starting price while R&B, Rock and Country have the lowest starting price.
4. Cities like Seattle, Pittsburgh and Melbourne have the highest starting price while Virginia, Springfield and New York have the lowest starting price.

5. From T-test and Bootstrapping sampling test, Mean minimum price of Ed Sheeran and G-Eazy were not the same but in fact the average minimum price of Ed Sheeran was greater than the average minimum price of G-Eazy.
6. From Chi-Square test, the Minimum price and Score of the Artists are dependent on each other.
7. From Anova tests, there is an effect of the Artist on the Minimum Price and a combination of City and Weekend also affect our target variable (minprice).
8. The artist of the event, venue of the event, population of the city the event is in and the rating of the artist, all are significant variables while predicting the Starting price of an event.
9. The dataset is raw and very random and there are of lot of other factors that may be required for us to better predict the starting price of any concert such as historic average prices of each venue, artist and so on.