Exploratory Data Analysis and Statistical Inference

Analytic Goals

1- Calculate descriptive statistics of data points/features that will be used in the recommender system.
2- Get a sense of the contents of the song dataset.
3- Gain a strong sense of the distribution of the data points/features that will be used in the recommender system. In particular, user listen counts, song listen counts and artist listen counts.
4- Test hypotheses about the user listen counts, song listen counts and artist listen counts means.
5- Explore relationship between user interactions and basic song features.
6- Cursory view of the top artist tags and tags to gain a sense of top genres represented in the dataset.

Major Findings in Exploratory Data Analysis and Statistical Inference

Song Metadata	**Descriptive Statistics**	905,531 unique songs
	Year of release pattern	Peak year release songs in 2000's.
	Artist song counts pattern	Right skewed with 3 peaks.
Song indicator (y=1, n=0) aggregated by: user, song and artist	**Descriptive Statistics**	1.019M unique users 42% of song the catalog listened 66.5% of artists listened
	Normality Testing: *Graphical and Hypothesis testing*	Highly skewed to the right. Distributions are **not normal**. Some distributions may not satisfy central limit theorem.
	One - sample mean hypothesis testing	High likelihood sample means are close to population means. **User Listen Indicator Means -** *User:* 47 | *Song:* 125 | *Artist:* 1,367
Relationships between song features and play indicator	**Joint plots with Pearson r**	No strong relationship between listens and song features
Artist Tags	**Top 20 artist tags from Echo Nest and MusicBrainz**	Top tags and terms: Rock, electronic and pop.

Part 1. Song Features Exploration

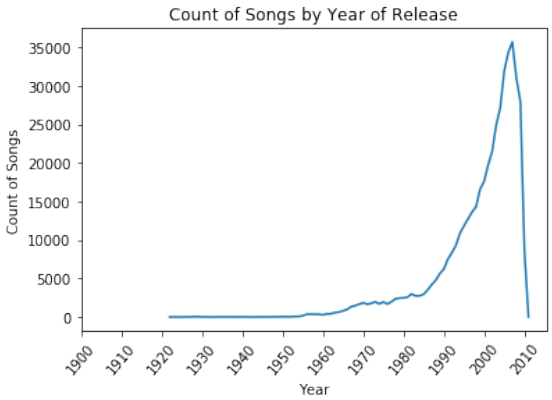
Basic Song Features Descriptive Statistics								
Song Feature	**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
artist_familiarity	905,531	0.55	0.14	0.00	0.48	0.56	0.64	1.00
artist_hotttnesss	905,700	0.38	0.12	0.00	0.33	0.38	0.44	1.08
song_hotttnesss	551,532	0.35	0.23	0.00	0.22	0.38	0.53	1.00
danceability	905,712	0.00	0.00	0.00	0.00	0.00	0.00	0.00
duration	905,712	246.91	124.88	0.31	179.93	227.79	286.41	3034.91
energy	905,712	0.00	0.00	0.00	0.00	0.00	0.00	0.00
key	905,712	5.31	3.59	0.00	2.00	5.00	9.00	11.00
loudness	905,712	-10.16	5.24	-58.18	-12.74	-8.99	-6.39	4.32
mode	905,712	0.67	0.47	0.00	0.00	1.00	1.00	1.00
tempo	905,712	123.84	35.28	0.00	97.73	121.69	144.90	302.30
time_signature	905,712	3.59	1.23	0.00	3.00	4.00	4.00	7.00
year	456,811	1999	10.29	1922	1995	2002	2006	2011

– **Finding #1:** Song Features Descriptive Statistics

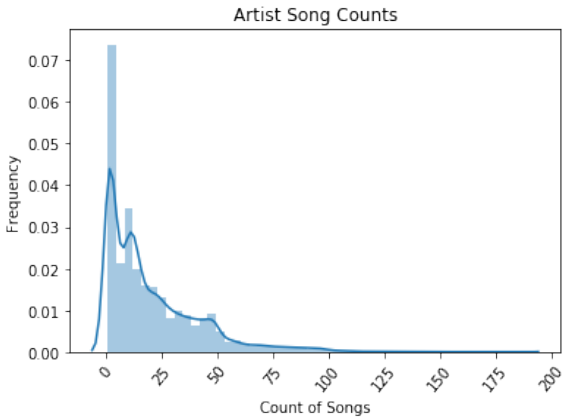
- Looking at the means it seems artist familiarity is 0.55, it falls somewhat in the middle of the possible values of 0 to 1.
- Artist hottness and song hottness mean values skew towards the lower range of hottness.
- In terms of song duration, the mean song is 4.10 minutes (which sounds accurate).
- The mean loudness of the tracks is -10.16, which suggests many of the tracks are not in the extreme of loudness (since this value is far away from max loudness of 4.31)
- The average key is 5.31 (this is around B major or D flat major).
- The mean beats per minute (tempo) is 123. Which means the mean song is in Allegro tempo: fast, quick, and bright.
- The mean mode is 0.66, there is an skew towards songs in the major scale.
- The mean time signature if 3.59 and the 75% percentile is 4. This means most songs are in 4/4 time signature.
- The mean year of release if 1998, with approximately. 25% of songs released between 2006 and 2011.

- **Finding #2:** There are 44,421 unique artists and 905,712 unique songs in the dataset.

- **Finding #3:** Year of release heavily skews to recent releases with a sharp peak in the 2000's. Year data is not complete, I have year data for 456,811 songs or 50% of the song dataset. It is unclear if the missing values are random or systematic, an example of this would be older songs tending to have missing values for year.



- **Finding #4:** Artist song counts are skewed to the right and have some peaks.



Part 2. User Listen Counts

Song Listens by User Descriptive Statistics								
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
play	1,019,318	47.46	57.82	10.00	16.00	27.00	55.00	4,400
play_count	1,019,318	136.05	184.53	10.00	34.00	73.00	163.00	13,132

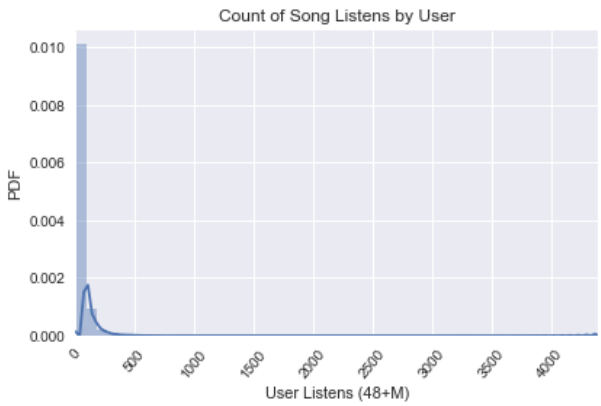
- **Finding #1:** There are 1,019,318 unique users.

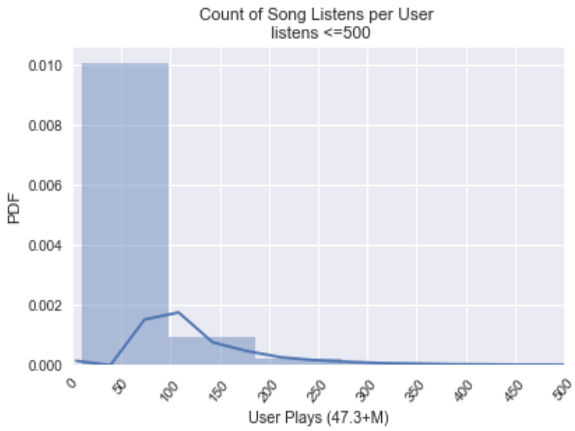
- **Finding #2:** The play indicator has a mean of 47.45 and a standard deviation of 57.82. This means that the average user has listened to 47.45 different songs.

- **Finding #3:** The percentile distribution suggests a long tail somewhere. The 75% percentile is 55 songs and the max song plays is 4,400 songs.

- **Finding #4:** The distribution of user song counts is heavily skewed to the right. The kurtosis and skewedness statistics suggest high levels of positive skewedness and kurtosis.

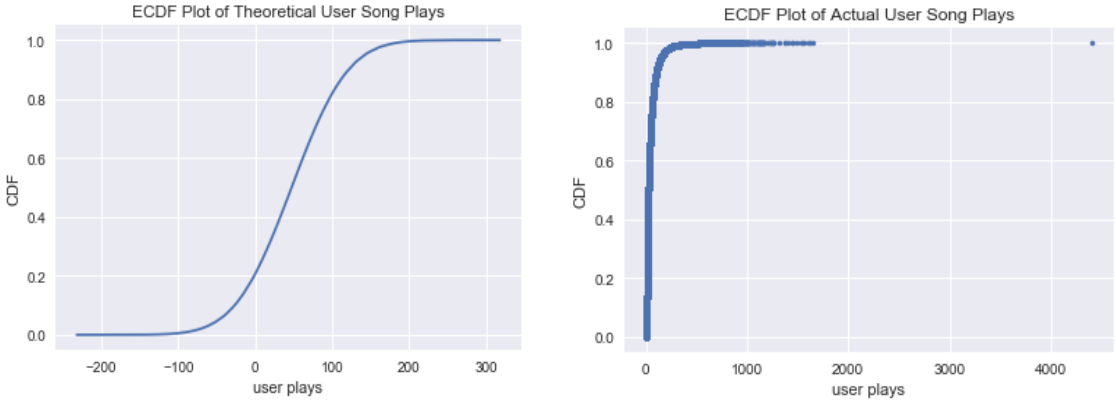
    - Skewness = 4.826
    - Kurtosis = 68.503





- **Finding #5:** Distribution of user listen is **not normally distributed**.

    - Graphical method shows that the user listen counts distribution vary significantly in shape from a theoretically normal distribution of the sample's mean and standard deviation.



    - Hypothesis testing method shows that distribution is **not normal**. The null hypothesis (H_0) is that the distribution is normal. The alternative hypothesis (H_a) is that is not normal. The testing is at the alpha level of 0.05. Since in both tests the p-value is less than 0.05 the null hypothesis is rejected and the alternative hypothesis is accepted that the listens by user distribution is not normal.

        - D'Agostino and Pearson's Normality Test[1]: stat=1059797.83, pvalue=0.0

        - Anderson-Darling Normality Test[2]: stat=105,800.29 > 0.05 critical value of 0.787

- **Finding #6:** The user listen count data satisfies the assumptions of the central limit theorem.

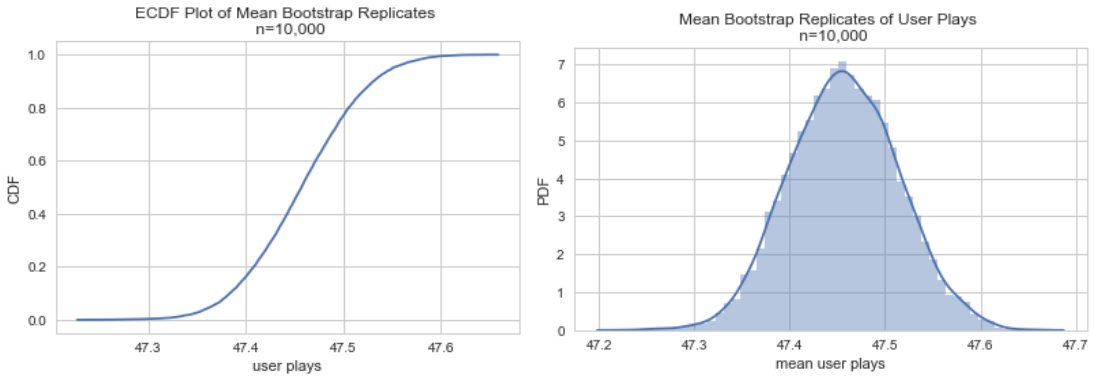
    - *Independence:* To satisfy this condition we will need to assume that the user's listen count is independent of the listen count of another user. This seems to me like a reasonable assumption. Nonetheless, it is unclear if this sample is 10% of the

[1] This function tests the null hypothesis that a sample comes from a normal distribution. It is based on D'Agostino and Pearson's, test that combines skew and kurtosis to produce an omnibus test of normality.
[2] The Anderson-Darling tests the null hypothesis that a sample is drawn from a population that follows a particular distribution. For the Anderson-Darling test, the critical values depend on which distribution is being tested against. This function works for normal, exponential, logistic, or Gumbel (Extreme Value Type I) distributions. This test has been shown to have more statistical power for testing normality.

population. The echo nest claims that this sample is a small subset of their music universe, how small is unclear. Find info here: http://blog.echonest.com/post/11992136676/taste-profiles-get-added-to-the-million-song

- *Randomness:* The Echo Nest randomly selected a sample of users whose play counts matched to the song ID's in the dataset.
- *Sample Size > 30*: The sample size is greater than 30.

- **Finding #8:** There is a 95% chance that the user listens population mean is between 47.34 - 47.57.
  - Using a bootstrap approach with 10,000 trials, the confidence interval for the user listen counts is 47.34 - 47.57. The mean replicates are normally distributed (see figures below).
  - This confidence interval is narrow and contains the sample mean of 47.45.



- **Finding #9:** There is a significant probability that the population mean is 45.47 listens. The null hypothesis (H_0) is that the population mean is 45.47 user song listens. The alternative hypothesis (H_a) is that the population mean is *not* 45.47 user song listens. The alpha level is 0.05. Since in both tests the p-value is greater than 0.05, the null hypothesis that the population mean is 45.47 user song listens is cannot be rejected.
  - Bootstrap hypothesis testing: pvalue = 0.4996
  - One sample t-test: stat=0.119, pvalue=0.905

User Listen Counts/Plays Conclusions

1. The distribution is not normal.
2. According to the bootstrap replicates of the mean and the t-test it is highly likely that the population user listen count mean is around 47.59 unique song listens per user.
3. There are several factors that could be driving these results and creating a heavily right skewed user listen count distribution:
    - The user play dataset is very large.
    - The data was collected in a certain timeframe and procedure and we do not have full user listening history.
    - The song data was extracted in December of 2011 which is holiday season, which could affect the user listen distribution (since only those who matched with the song dataset are included).
    - There are specific characteristics of users with high user-song interactions (such as business customer, several people in the same account etc.).
    - The structure of the music industry, in which a few songs/artists create hits and most working artists do not.

Part 3. User Listen Counts per Song

User Listens by Songs Descriptive Statistics								
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
play	384,546	125.79	799.03	1.00	4.00	13.00	52.00	110,479
play_count	384,546	360.63	3,256.81	1.00	8.00	32.00	133.00	726,885

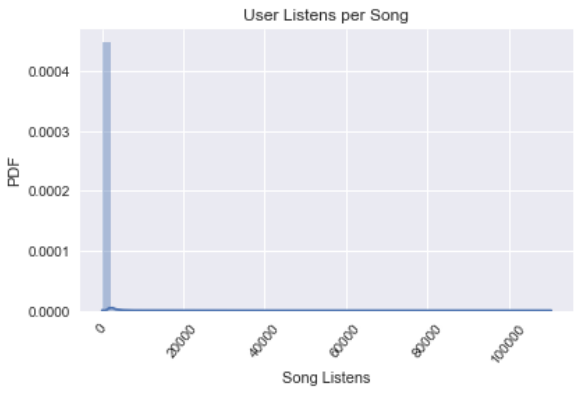
- **Finding #1:** There are 384,546 unique songs that have been listened to. This means that 42% of the song catalog has been listened to and 58% has not been listened to at all.

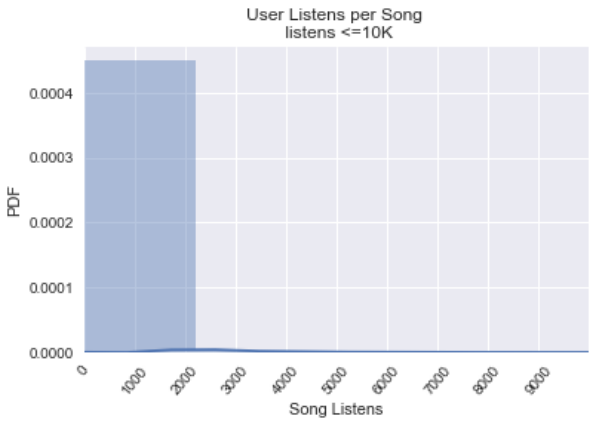
- **Finding #2:** As shown in the previous section, at most a user as listed to 4,400 unique songs, and the average user has listened to 47.45 different songs. We can conclude that the song catalog and user listens are sparsely distributed (most users have listened to a small portion of the available catalog).

- **Finding #3:** The play indicator has a mean of 125.79 and a standard deviation of 799. This means that the average song has 125.79 unique user listens.

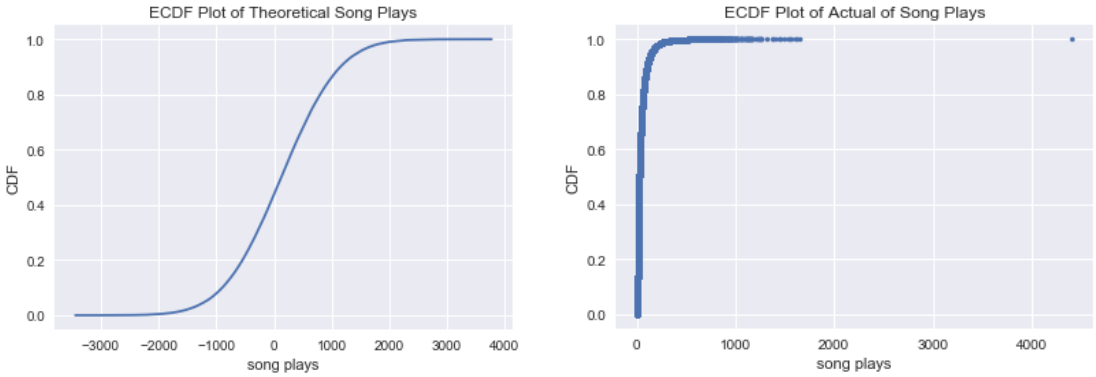
- **Finding #4:** The percentile distribution suggests a long tail somewhere. The 75% percentile is 52 songs and the max song plays is 110,479 unique user listens.

    - Skewness = 46.11
    - Kurtosis = 3780.29





- **Finding #5:** Distribution of user listens per song is **not normally distributed**.



  - Graphical method shows that the user song listens per track distribution vary significantly in shape from a theoretically normal distribution of the sample's mean and standard deviation.

  - Hypothesis testing method shows that distribution is **not normal**. The null hypothesis (H_0) is that the distribution is normal. The alternative hypothesis (H_a) is that is not normal. The testing is at the alpha level of 0.05. Since in both tests the p-value is less than 0.05, the null hypothesis is rejected and the alternative hypothesis is accepted that the listens by user distribution is not normal.

    - D'Agostino and Pearson's Normality Test[3]: stat=1176273.18, pvalue=0.0

    - Anderson-Darling Normality Test[4]: stat= 110035.21 > 0.05 critical value of 0.787

- **Finding #6:** It is suspect that the distribution satisfies the conditions of the central limit theorem.
  - *Independence:* To satisfy this condition we will need to assume that a song play count is independent of the play count of another song. In this case, I think this may

[3] This function tests the null hypothesis that a sample comes from a normal distribution. It is based on D'Agostino and Pearson's, test that combines skew and kurtosis to produce an omnibus test of normality.
[4] The Anderson-Darling tests the null hypothesis that a sample is drawn from a population that follows a particular distribution. For the Anderson-Darling test, the critical values depend on which distribution is being tested against. This function works for normal, exponential, logistic, or Gumbel (Extreme Value Type I) distributions. This test has been shown to have more statistical power for testing normality.

not be not a reasonable assumption to make due to popularity, artists similarity, song similarity and how the echo nest service decided to show the catalog to their users. Also, it is unclear if this sample is 10% of the population.

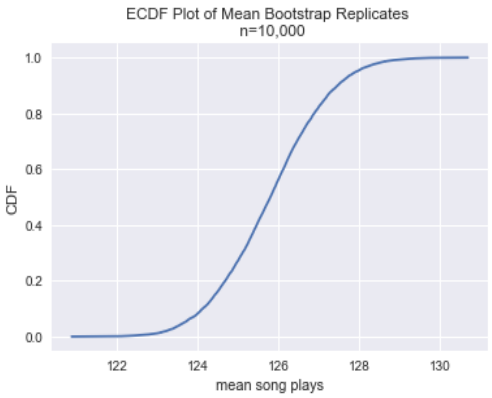
- *Randomness:* The Echo Nest randomly selected a sample of users whose play counts matched to the song ID's in the dataset.

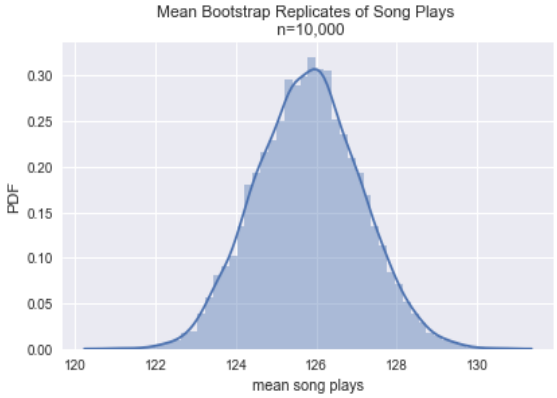
- *Sample Size > 30:* The sample size is greater than 30.

– **Finding #8:** There is a 95% chance that the user listens population mean is between 47.34 - 47.57.

  - Using a bootstrap approach with 10,000 trials, the confidence interval for the user listens per song is 123.31- 128.34. The mean replicates are normally distributed (see figures below).

  - This confidence interval contains the sample mean of 125.79.





– **Finding #9:** There is a significant probability that the population mean is 45.47 listens. The null hypothesis (H_0) is that the population mean is 45.47 user song listens. The alternative hypothesis (H_a) is that the population mean is *not* 45.47 user song listens. The alpha level of 0.05. Since in both tests the p-value is greater than 0.05, the null hypothesis that the population mean is 125.79 user song listens cannot be rejected. The t-test should be taken with some suspicion since this distribution may not satisfy the central limit theorem.

  - Bootstrap hypothesis testing: pvalue = 0.496

  - One sample t-test: stat=0.003, pvalue= 0.997

User Listens/Plays per Song Conclusions

1- The distribution is not normal.
2- According to the bootstrap replicates of the mean and the t-test it is highly likely that the population user listen count by song mean is around 125.79. The t-test should be taken with some suspicion since this distribution may not satisfy the central limit theorem.
3- There are several factors that could be driving these results and creating a heavily right skewed user listen count distribution:
    a. The user play dataset is very large.
    b. The data was in a certain timeframe and procedure and we do not have full user listening history.
    c. The song data was extracted in December of 2011 which is holiday season, which could affect the user song play behavior.
    d. There are specific characteristics of the user with high user play interactions (such as business customer, several people in the same account etc.).
    e. The structure of the music industry, in which a few songs/artists create hits and most working artists do not.

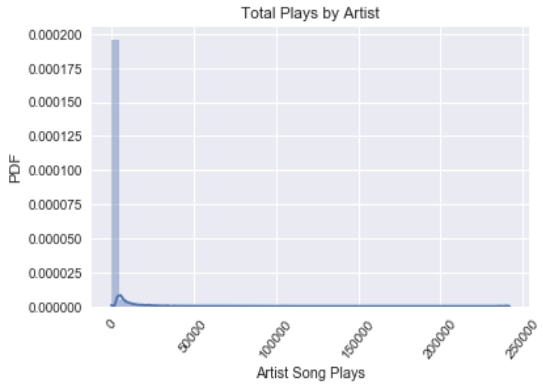
Part 4. User Listen Counts per Artist

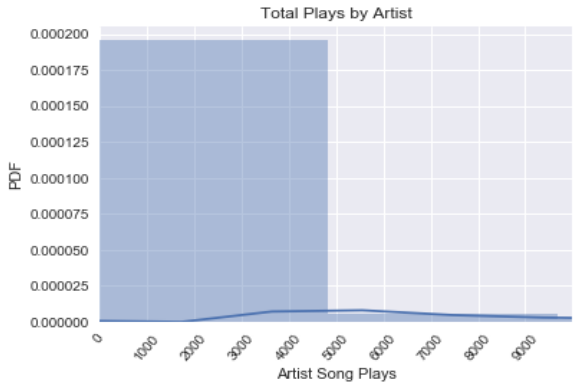
User Listens by Artist Descriptive Statistics								
	count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
play	29,559.0	1,367.44	6,498.71	1.00	20.00	112.00	569.50	241,823.00
play_count	29,559.0	3,895.82	18,704.71	1.00	50.00	323.00	1,699.00	884,464.00

- **Finding #1:** There are 29,559 unique artists that have been listened to. This means that 66.5% of artists in the song dataset has been listened to, and 33.5% artists have not been listened to at all.

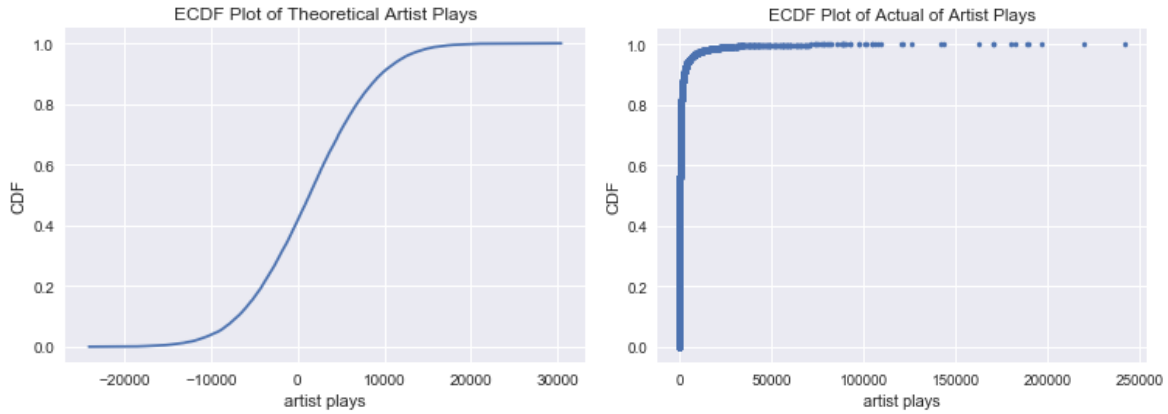
- **Finding #2:** The user listen indicator by artist has a mean of 1,367.44 and a standard deviation of 6,498.71. This means that the average artist has 1,367.44 unique user listens.

- **Finding #3:** The percentile distribution suggests a long tail somewhere. The 75% percentile is 569.50 songs and the max song plays is 241,823 unique user listens.





- **Finding #4:** Distribution of user listens per artist is **not normally distributed**.
    - Skewness = 15.13
    - Kurtosis = 344.37



- Graphical method shows that the user listens by artist distribution vary significantly in shape from a theoretically normal distribution of the sample's mean and standard deviation.

- Hypothesis testing method shows that distribution is **not normal**. The null hypothesis (H_0) is that the distribution is normal. The alternative hypothesis (H_a) is that is not normal. The testing is at the alpha level of 0.05. Since in both tests the p-value is less than 0.05, the null hypothesis is rejected and the alternative hypothesis is accepted that the listens by user distribution is not normal.

  - D'Agostino and Pearson's Normality Test[5]: stat= 56969.63, pvalue=0.0

  - Anderson-Darling Normality Test[6]: stat= 8001.32 > 0.05 critical value of 0.787

– **Finding #6:** It is suspect that the distribution satisfies the conditions of the central limit theorem.

  - *Independence:* To satisfy this condition we will need to assume that an artist play count is independent of the play count of another artist play count. In this case, I think this may not be not a reasonable assumption to make due to popularity effects, artists similarity, song similarity and how the echo nest service decided to show the catalog to their users. However, it may be a reasonable assumption that this sample represents <10% of the artist population.

[5] This function tests the null hypothesis that a sample comes from a normal distribution. It is based on D'Agostino and Pearson's, test that combines skew and kurtosis to produce an omnibus test of normality.
[6] The Anderson-Darling tests the null hypothesis that a sample is drawn from a population that follows a particular distribution. For the Anderson-Darling test, the critical values depend on which distribution is being tested against. This function works for normal, exponential, logistic, or Gumbel (Extreme Value Type I) distributions. This test has been shown to have more statistical power for testing normality.

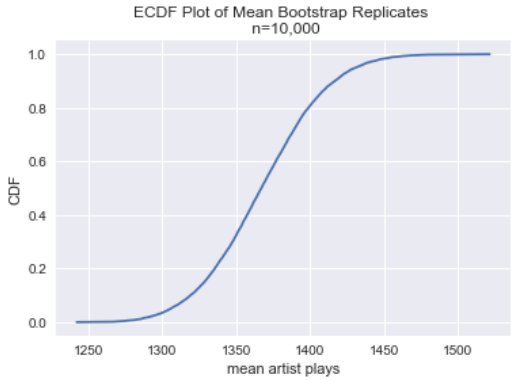
- *Randomness:* The Echo Nest randomly selected a sample of users whose play counts matched to the song ID's in the dataset.

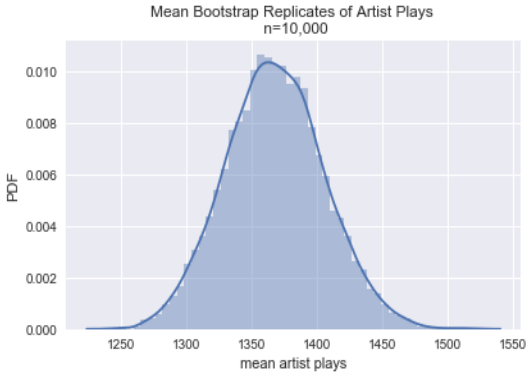
- *Sample Size > 30: The sample size is greater than 30.*

- **Finding #7:** There is a 95% chance that the artists listens population mean is between 1294.93- 1443.78.

    - Using a bootstrap approach with 10,000 trials, the confidence interval for the user listens by artist is 1294.93- 1443.78. The mean replicates are normally distributed (see figures below).

    - This confidence interval contains the sample mean of 1,367.44.





- **Finding #9:** There is a significant probability that the population mean is 1,367.44 user listens per artist. The null hypothesis (H_0) is that the population mean is 1,367.44 user listens per artist. The alternative hypothesis (H_a) is that the population mean is *not* 1,367.44 user listens per artist. The alpha level of 0.05. Since in both tests the p-value is greater 0.05, the null hypothesis that the population mean is not 1,367.44 user listens per artists listens is cannot be rejected. The t-test should be taken with some suspicion since this distribution may not satisfy the central limit theorem.

    - Bootstrap hypothesis testing: pvalue = 0.5033

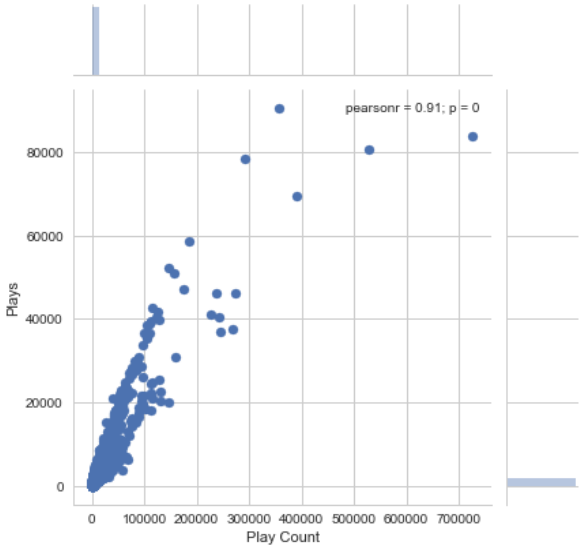
    - One sample t-test: stat= -4.202, pvalue= 0.999

User Listens per Artist Conclusions

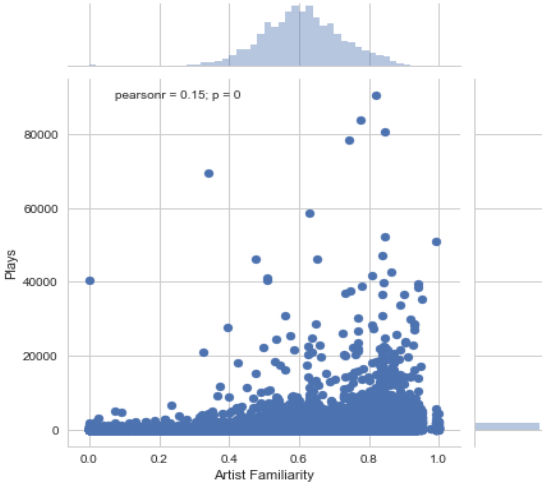
1- The distribution is not normal.
2- According to the bootstrap replicates of the mean and the t-test it is highly likely that the population user listen count per artist mean is around 1,367.44. The t-test should be taken with some suspicion since this distribution may not satisfy the central limit theorem.
3- There are several factors that could be driving these results and creating a heavily right skewed user listen per artist distribution:
    a. The user play dataset is very large.
    b. The data was in a certain timeframe and procedure and we do not have full user listening history.
    c. The song data was extracted in December of 2011 which is holiday season, which could affect the user song play behavior.
    d. There are specific characteristics of the user with high user play interactions (such as business customer, several people in the same account etc.).
    e. The structure of the music industry, in which a few songs/artists create hits and most working artists do not.

Part 5. Relationships between basic song features and song listens

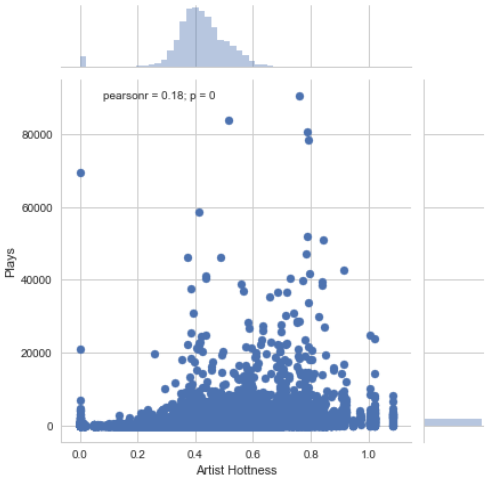
- There does not seem to be a strong relationship between unique song plays/interactions and any of the basic song features.
- The lack of relationship may be due to the sparsity of the data. For example, most users having not listened to that much of the catalog to result in meaningful patterns.
- The lack of relationship may also be due to the fact that many of the songs may be too heterogeneous or different.
- They may be patterns or correlations if features are looked at in combination rather than isolated.
- There seems to be a strong linear relationship between play count and plays. This means that there is a tendency in that the higher the play counts, the higher the number of unique user plays. This makes sense and ought to be expected given that the play indicator is derived from the play count.
- The high person correlation coefficient value of the relationship between user plays and user play counts leads me to think that song/artist popularity maybe an important factor in song/artist/user interactions (i.e. the more people hear about a song/artist the more likely they are to interact with it and have play it more).



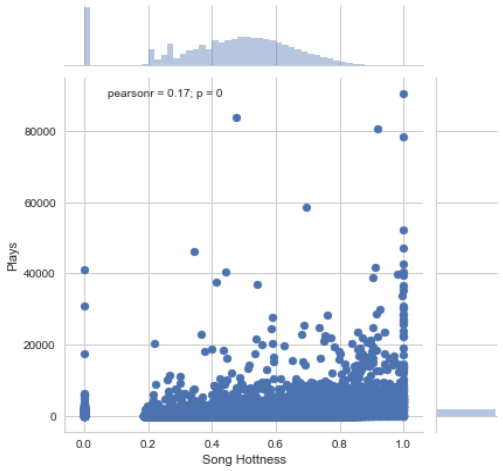
Play Count and Plays



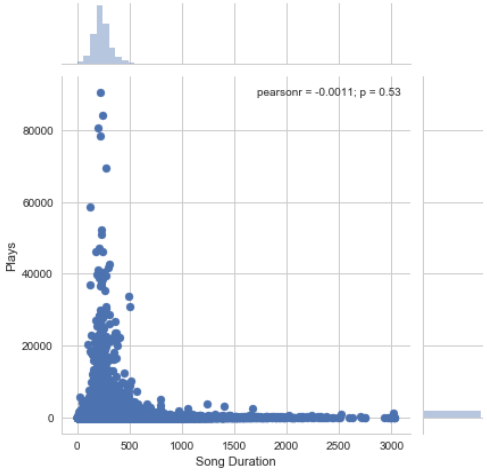
Artist Familiarity



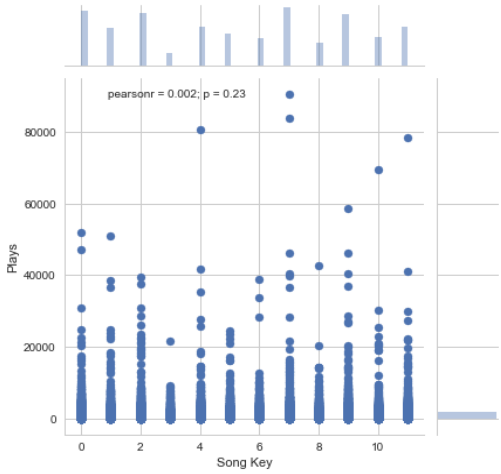
Artist Hotness



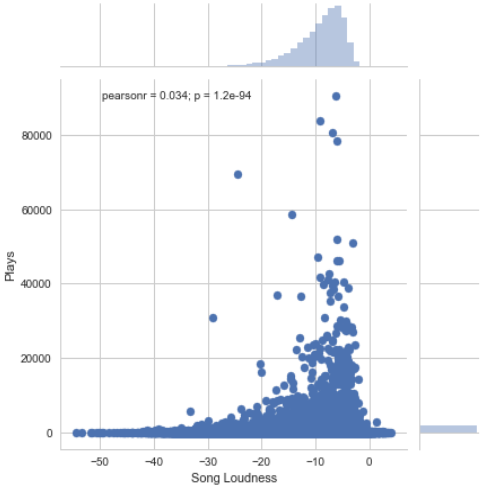
Song Hotness



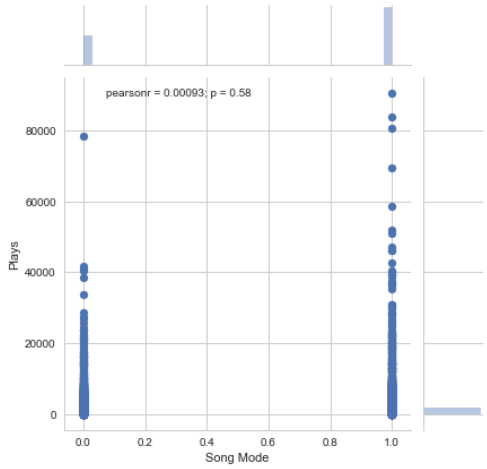
Song Duration



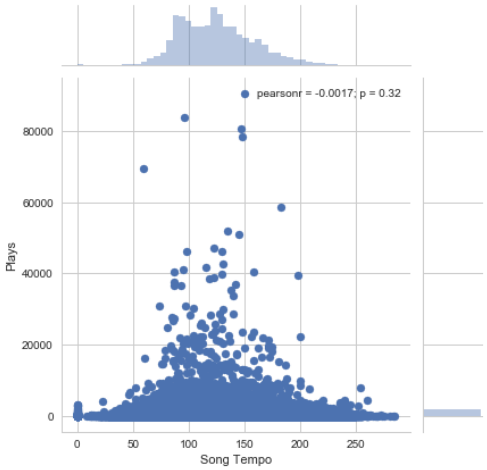
Song Key



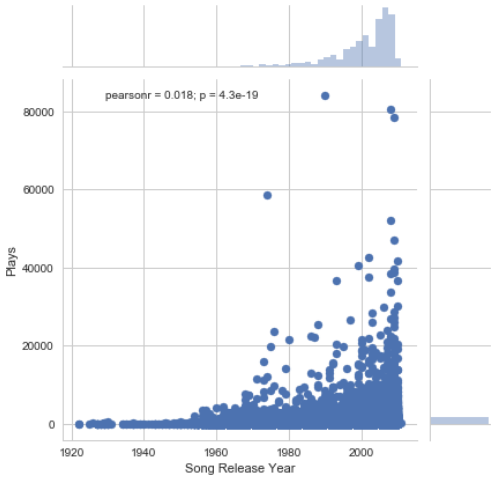
Song Loudness



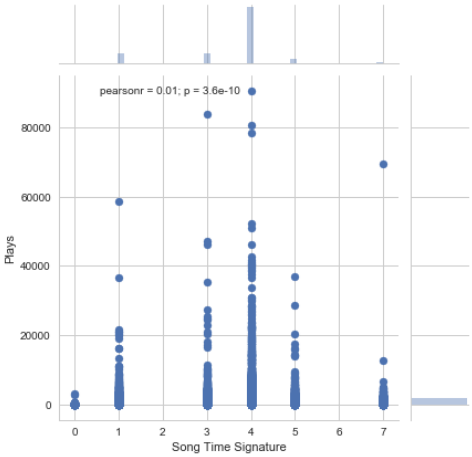
Song Mode



Song Tempo

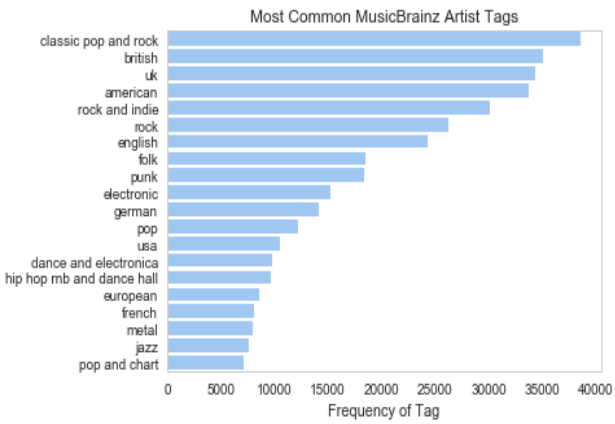


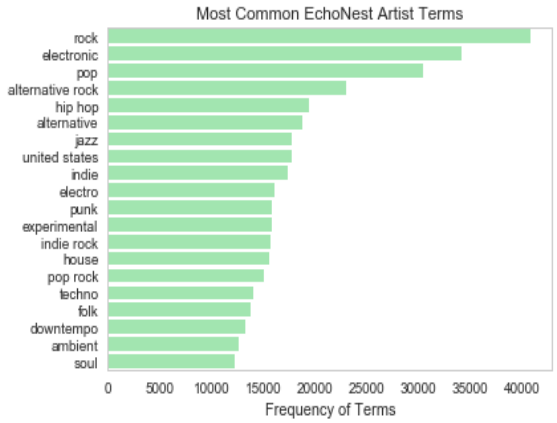
Song Release Year



Time Signature

Part 6. Artist Tags and Terms





- **MusicBrainz Tags**

    - There is a total of 2,321 unique MusicBrainz tags.
    - The average tag has ~344 unique artists associated with it.
    - There seems to be skewedness, 75% of the tags have 104 artists while the maximum artists associated with a tag is ~38.5K.
    - As discussed previously there are 44,421 unique artists in the song dataset. So this means that 98% of artists in the MSD dataset has one or more MusicBrainz tag.
    - Classic pop, rock, british, uk, american indie, english and folk are the most popular artists tags in the MusicBrainz data.

- **EchoNest Terms**

    - There is a total of 7,643 unique terms.
    - The average tag has ~212 unique artists associated with it.
    - There seems to be some skewedness, 75% of the tags have 32 unique artists while the maximum artists associated with a tag is ~40.8K.
    - As discussed previously, there are 44,421 unique artists in the song dataset. unique artists in the dataset. So this means that 98% of artists in the MSD dataset has one or more EchoNest tag.
    - The Echo Nest has almost 3x more unique tags than the Music Brainz array.
    - Rock, electronic, pop, hip hop, alternative and jazz the most popular artists terms in the Echo Nest data. Rock intersects with the top tags of the MusicBrainz.