Data Appendix

Analysis Data File 1

- Unit of Observation: Each row in the dataset represents an individual song.
- Total Observations: 343,195
- Total Observations after cleaning and transformations: 9,382

Variables

1. Variable name: chart_date
   - Type: object to datetime
   - Description: The date that the Billboard Hot 100 chart was released.
   - Observations: 343,195(343,195)
   - Observations after transformations: 9,382(9,382)
   - Transformations: The variable was converted into a datetime object and sorted into ascending to descending order. Removed rows made before the year 2005 as streaming took over the majority of music industry sales starting in 2016, streamlining our dataset to see how streaming influenced song lengths by having 11 years prior to this changepoint. All observations except for the highest instance (the amount of time a song has returned to the chart after more than 1 week off the chart) value were removed to eliminate duplicate entries for songs, influencing the overall observations for this variable as well.
   - Frequency table grouped by top 10 years after transformations:
     -

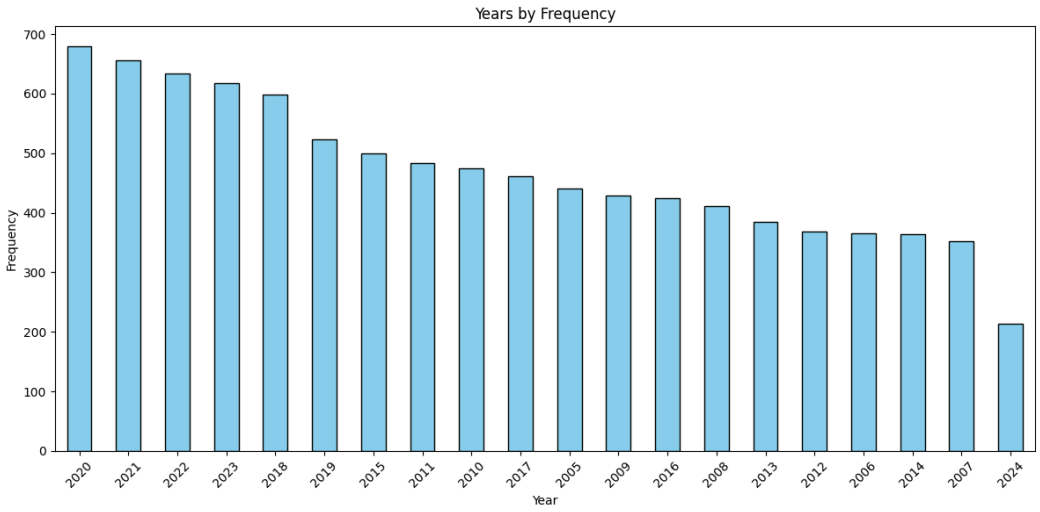
chart_date	Count
2020	679
2021	656
2022	634
2023	618
2018	599
2019	523
2015	500
2011	484
2010	474

chart_date	Count
2020	679
2021	656
2022	634
2023	618
2018	599
2019	523
2015	500
2011	484
2017	462

- Summary statistics after transformations:
  -

count	9382
mean	2015-08-23
min	2005-01-01
25%	2010-10-16
50%	2016-04-23
75%	2020-08-27
max	2024-04-27

- Bar chart of frequency distribution:

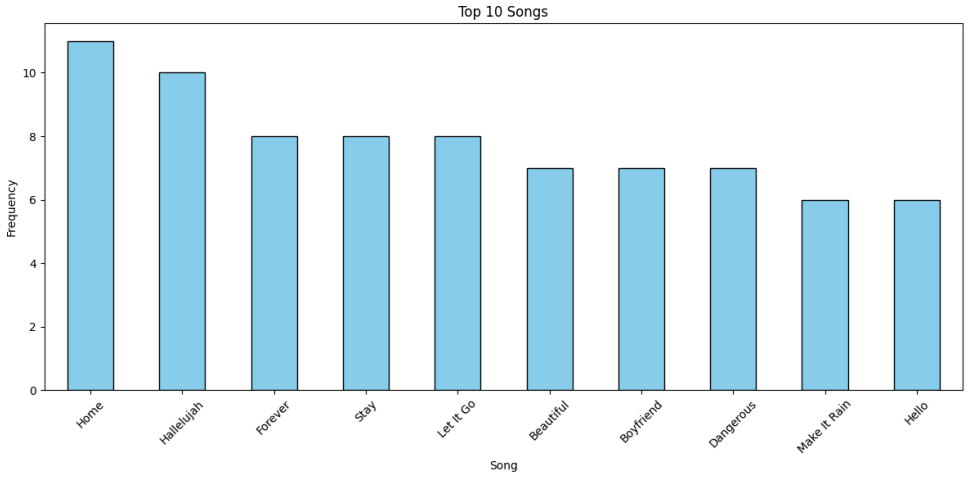


---

2. Variable name: song
   - Type: object
   - Description: The name of the song
   - Observations: 343,195(343,195)
   - Observations after transformations: 9,382(9,382)
   - Transformations: All observations except for the highest instance (the amount of time a song has returned to the chart after more than 1 week off the chart) value were removed to eliminate duplicate entries for songs, influencing the overall observations for this variable as well.
   - Frequency table of top 10 songs after date transformations:

song	Frequency
Home	11
Hallelujah	10
Forever	8
Stay	8
Let It Go	8
Beautiful	7
Boyfriend	7
Dangerous	7
Make It Rain	6

Hello	6

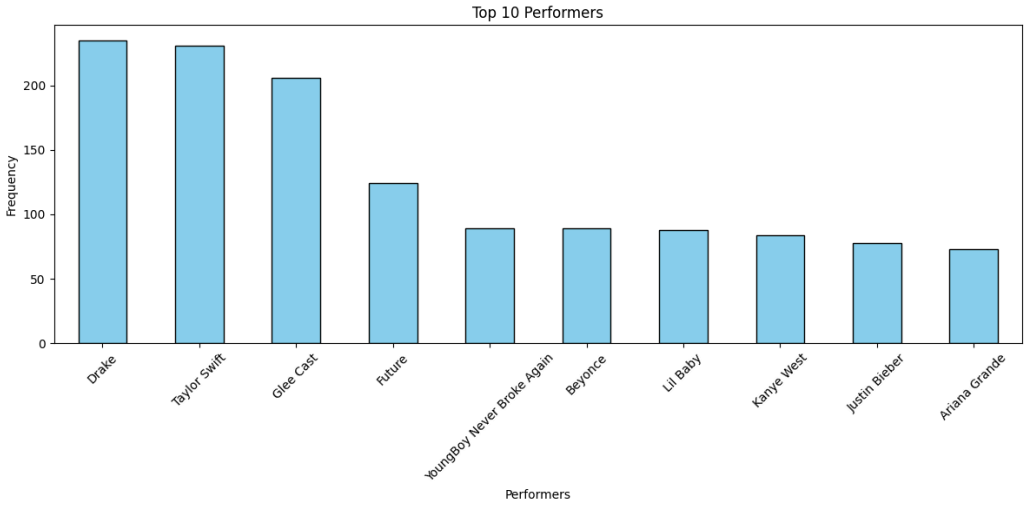
- Bar chart showing frequency distribution of top 10 songs:
  - 

---

3. Variable name: performer
   - Type: object
   - Description: The performer of the song
   - Original observations: 343,195(343,195)
   - Observations after transformations: 9,382(9,382)
   - Transformations: Cleaned this variable to only include the first artist by removing the featuring artists. This allowed our Spotify API to successfully run. All observations except for the highest instance (the amount of time a song has returned to the chart after more than 1 week off the chart) value were removed to eliminate duplicate entries for songs, influencing the overall observations for this variable as well.
   - Frequency table of top 10 artists after transformations:
     -

Performer	Frequency
Drake	235
Taylor Swift	231
Glee Cast	206
Future	124
YoungBoy Never Broke Again	89
Beyonce	89
Lil Baby	88

Kanye West	84
Justin Bieber	78
Ariana Grande	73

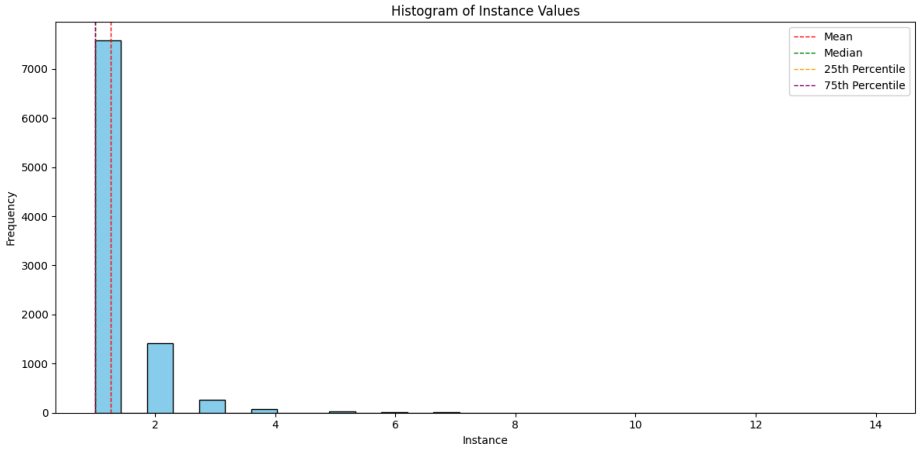
- Bar chart showing frequency distribution of top 10 artists:
  - 

---

4. Variable name: instance
   - Type: float64
   - Description: Indicates how many times a song has returned to the chart after more than 1 week off the chart
   - Observations: 343,195(343,195)
   - Observations after transformations: 9,382(9,382)
   - Transformations: All observations except for the highest instance (the amount of time a song has returned to the chart after more than 1 week off the chart) value were removed to eliminate duplicate entries for songs. This helps streamline the data for API processing, reducing wait times caused by multiple instances of the same song on the chart.
   - Summary statistics after date transformations:
     -

count	9,382
mean	1.266
std	0.7
min	1.0
25%	1.0

50%	1.0
75%	1.0
max	14

- Histogram of Summary Statistics:
  - 

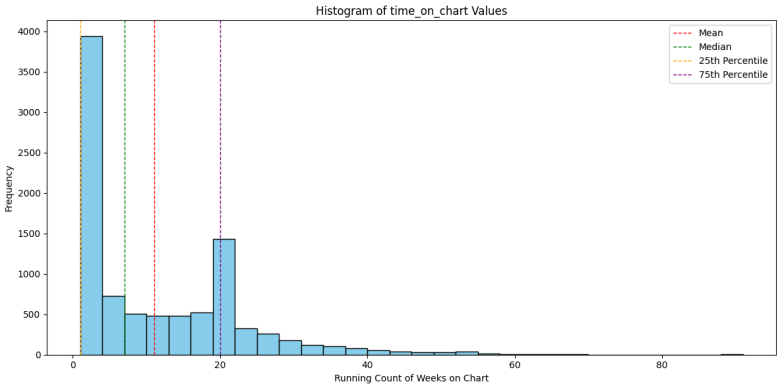
---

5. Variable name: time_on_chart
   - Type: int64
   - Description: The running count of weeks (all-time) a song has been on the chart
   - Observations: 343,195(343,195)
   - Observations after transformations: 9,382(9,382)
   - Transformations: All observations except for the highest instance (the amount of time a song has returned to the chart after more than 1 week off the chart) value were removed to eliminate duplicate entries for songs, influencing the overall observations for this variable as well.
   - Summary statistics after transformations:
     -

count	9,382
mean	11.1
std	11.45
min	1.0
25%	1.0
50%	7.0
75%	20.0

max	91.0

- Histogram of Summary Statistics:

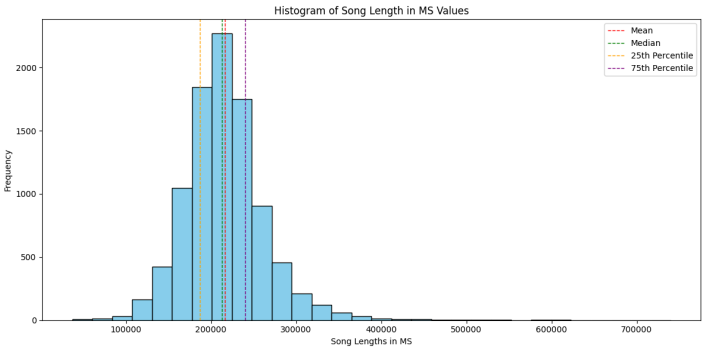


---

6. Variable name: song_length_ms
   - Type: int64
   - Description: Length of songs in milliseconds
   - Observations: 9,382(9,382)
   - Transformations: This variable was created through running a LastFM and Spotify API to provide song length data.
   - Summary statistics after transformations:

count	9,382
mean	216049.65
std	48044.61
min	37013
25%	187108
50%	212949.5
75%	239809.75
max	740010

   - Histogram of Summary Statistics:

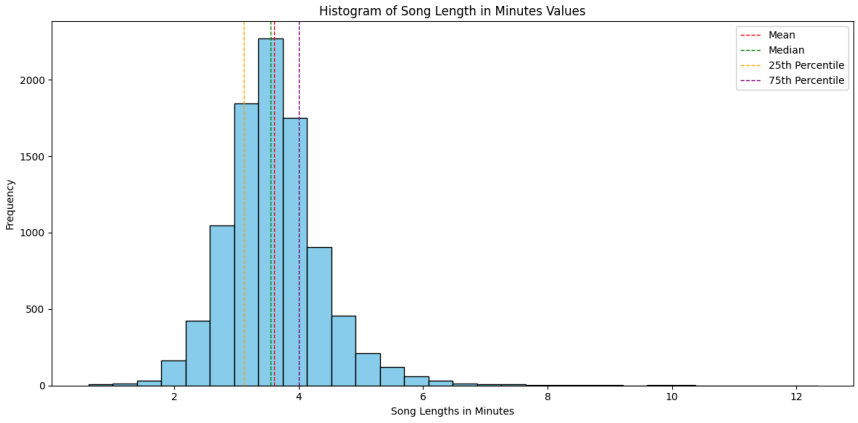


---

7. Variable name: song_length_mins
   - Type: float64
   - Description: Length of songs in minutes
   - Observations: 5,309(5,309)
   - Transformations: This variable was created through running a LastFM and Spotify API to provide song length data.
   - Summary statistics after transformations:

count	9,382
mean	3.6
std	0.8
min	0.62
25%	3.12
50%	3.55
75%	3.99
max	12.33

   - Histogram of Summary Statistics:



○

---

8. Variable name: chart_position
   - Type: int64
   - Description: The position of the song for the given chart date
   - Observations: 343,195(343,195)
   - Transformations: This variable was dropped from the dataset as it was unnecessary in answering our hypothesis.

---

9. Variable name: song_id
   - Type: object
   - Description: A concatenation of a song and performer to create a unique identifier
   - Observations: 343,195(343,195)
   - Transformations: This variable was dropped from the dataset as it was unnecessary in answering our hypothesis.

---

10. Variable name: consecutive_weeks
    - Type: float64
    - Description: For the given instance, how many weeks has the song been on the chart consecutively. A null value indicates the start of a new instance.
    - Observations: 343,195(343,195)
    - Transformations: This variable was dropped from the dataset as it was unnecessary in answering our hypothesis.

---

11. Variable name: previous_week
    - Type: float64
    - Description: For the given instance, what was the chart_position for the previous week
    - Observations: 343,195(343,195)
    - Transformations: This variable was dropped from the dataset as it was unnecessary in answering our hypothesis.

---

12. Variable name: peak_position
    - Type: int64
    - Description: Indicated the all-time best/peak position for a song_id
    - Observations: 343,195(343,195)
    - Transformations: This variable was dropped from the dataset as it was unnecessary in answering our hypothesis.

---

13. Variable name: worst_position
    - Type: float64
    - Description: Indicated the all time worst/lowest position for a song_id
    - Observations: 343,195(343,195)
    - Transformations: This variable was dropped from the dataset as it was unnecessary in answering our hypothesis.

---

14. Variable name: chart_debut
    - Type: object
    - Description: The date of the first initial instance for a song_id
    - Observations: 343,195(343,195)
    - Transformations: This variable was dropped from the dataset as it was unnecessary in answering our hypothesis.

---

15. Variable name: chart_url
    - Type: object
    - Description: This URL takes you to the chart on BillBoard.com
    - Observations: 343,195(343,195)
    - Transformations: This variable was dropped from the dataset as it was unnecessary in answering our hypothesis.