IS6030
Data Management
HW5 Submission

Krishna Teja Jagarlapudi
M10896584

TABLE OF CONTENTS

DESCRIPTION OF DATA

PRELIMINARY INTRODUCTION

- The data pertains to rock music played on multiple channels and on multiple days.
- Preliminary analysis of data –
    - 7 days, starting 16th June 2014 to 22nd June 2014
    - 25 radio stations
    - 384 artists
    - 1612 songs
- I obtained this data from the GitHub profile of fivethirtyeight.com by Nate Silver.
- The actual URL is - https://github.com/fivethirtyeight/data/tree/master/classic-rock

DESCRIPTION OF THE FILES

There are two .csv files -

- *classic-rock-raw-data.csv* – Each line represents a play of a song on a radio station. This data is obtained by scraping online sources and has 37673 rows and 9 columns:
    - RAW_SONG – the name of the song as scraped by the data provider
    - Song clean – Cleaned up version of above column
    - RAW_ARTIST - the name of the artist as scraped by the data provider
    - Artist clean – Cleaned up version of above column
    - Call sign – Call sign of the station that is playing the song. This is basically a codified version of the station name
    - Time - Time the song was Played. It is recorded as seconds since January 1, 1970
    - PlayID – Unique ID for each song play.
    - Combined – A combination of song and artist in the form – "<song> by <artist>"
    - First? – a dummy column
- *classic-rock-song-list.csv* - Each line represents one song in the set. There are 2230 rows and 5 useful columns in this file. The remaining three are dummy columns which are not pertinent for analysis.
    - Song clean – same as in above file
    - Artist clean – same as in above file
    - Release Year – Year in which the song was released
    - Combined – same as in above file
    - Play count – The number of times the song is played.
- The 'combined' column can be used to compare the data from the two tables.

NORMALIZATION OF DATA

- The data in *classic-rock-raw-data.csv* is of the first normal form. There is one row for every time a song is played on a radio station.

ISSUES WITH THE DATA

- The column names in the data have spaces in them. Some column names were also sql keywords.
    - The columns have been manually renamed.
- There are rows in *classic-rock-raw-data.csv* which have null values in some columns –
    - Null in song column leads to some artists having zero songs but non-zero plays
    - Null in artists column for a few plays but not in few others

  Such rows have been removed.

- Some songs have recording error – with release date as 1071 etc. To remove such errors, songs released before 1950 have been removed.
- Hence, though the second file has information about 2k+ songs, only 1612 songs are in the analysis.

SUMMARY OF ANALYSES

- Have used three tools – SQL, Tableau, and SAS - to perform manipulation and analysis activities on the data.
- It is seen that on average, each day, a radio station plays 5377 songs, with 1178 unique songs being played. This leads to each song played an average of 4 times a day.
- Led Zeppelin, Rolling stones, Van Halen, Pink Floyd and Tom Petty & the Heartbreakers are the top five most played bands.
- This corroborates with the external knowledge of rock music – these five bands are indeed some of the legends in rock music.
- Some of the one-hit-wonders (artists who have no more than 2 songs in the set) are Thin Lizzy and Derek and the Dominoes. Here too this makes sense.
- The top played song is 'Dream On' by Aerosmith. It was played 142 times in the one week period, with 29 plays on Wednesday, 19th June 2014.
- It is seen that, on Tuesday, all radio stations play lesser number of songs. The average plays is around 3.5k, compared to 5.3k of overall average plays. This is a statistically significant difference in the number of song plays with respect to the other days.
- **Suggestion:** One of two scenarios are happening
    - The radio stations have too many advertisements on Tuesdays. This will affect listener loyalty, who will soon catch on and stop listening on Tuesdays. The spread of advertisements must be evened out.
    - Radio stations noticed that users are not listening much to music on Tuesdays and so are playing lesser music.
    - The reason behind this is currently unknown. A dedicated market research team must be setup to analyze this issue. Instead of reacting to market demand, measures can be taken to create it.

MANIPULATION AND ANALYSIS OF DATA USING SQL

MANIPULATION OF DATA

- Importing data:
  - The two files are *.csv* type. It is easier to import *.xls* type files into MS SQL server studio.
  - Manually saved the *.csv* files as *.xls* files using MS Excel, and imported these files using the import data wizard.
  - Changed the column names manually to easier to use format.
- Fixing the air date:
  - The time at which the song was aired is represented as seconds from Jan 1st 1970.
  - Created a new column and calculated air date from the above information using `alter` and `update` queries.
- Combining the two datasets:
  - Created a new table by joining the two tables, on the `'combined'` column.
  - Only the `'Release Year'` column is taken from the second table; and `'play ID', 'Song clean', 'Artist clean', 'Combined', 'Call sign'` and `'air date'` are taken from the first.
  - The new table, named `'rock'`, is used for the rest of the analysis.

BASIC STATISTICS

- The below results from SQL speak of some basic information about the data:
  - Note that the average artist has 4 songs in the set.
  - On average, the songs in the set are from around 1986.

	Num_stations	Num_artists	Num_songs	Songs_per_artist	Avg_rel_year
1	25	384	1612	4	1986

- Average activity stats of the radio stations per day – number of song plays, number of unique songs, number of unique artists, number of repeat plays for a song, and number of songs per artist.

	Avg_num_plays	Avg_num_unique_songs	Avg_num_unique_artists	Avg_plays_per_song	Avg_plays_per_artist
1	5377	1178	302	4	17

TOP CHARTS

- Top artists in the set.
  - This is determined by looking at total number of song plays across all the artists' songs.
  - A filter that there must at least be 6 songs (1.5 times the average) from the artists is also used to take care of one-hit-wonders.
  - The age column is calculated as the average difference between the release year and 2014, the year it is being aired in.

	Artist	Num_plays	Age
1	Led Zeppelin	1546	41.82
2	Rolling Stones	1112	42.32
3	Van Halen	1055	32.89
4	Pink Floyd	1044	37.29
5	Tom Petty & The Heartbreakers	916	29.57

- One hit wonders.
    - This is determined by looking at top played artists with less than 2 songs (half the average) in the set.

	Artist	Num_plays	Age
1	Cheap Trick	167	36.66
2	Thin Lizzy	122	38.00
3	Steppenwolf	118	46.00
4	Manfred Mann	116	38.21
5	Derek & The Dominos	103	44.00

- Top played song on each date.

	Date_Aired	Song	Artist	Listens
1	2014-06-16	All Along the Watchtower	Jimi Hendrix	24
2	2014-06-17	Bohemian Rhapsody	Queen	20
3	2014-06-18	Dream On	Aerosmith	29
4	2014-06-19	Crazy On You	Heart	24
5	2014-06-20	Carry On Wayward Son	Kansas	27
6	2014-06-21	You Shook Me All Night Long	AC/DC	24
7	2014-06-22	Dance the Night Away	Van Halen	24

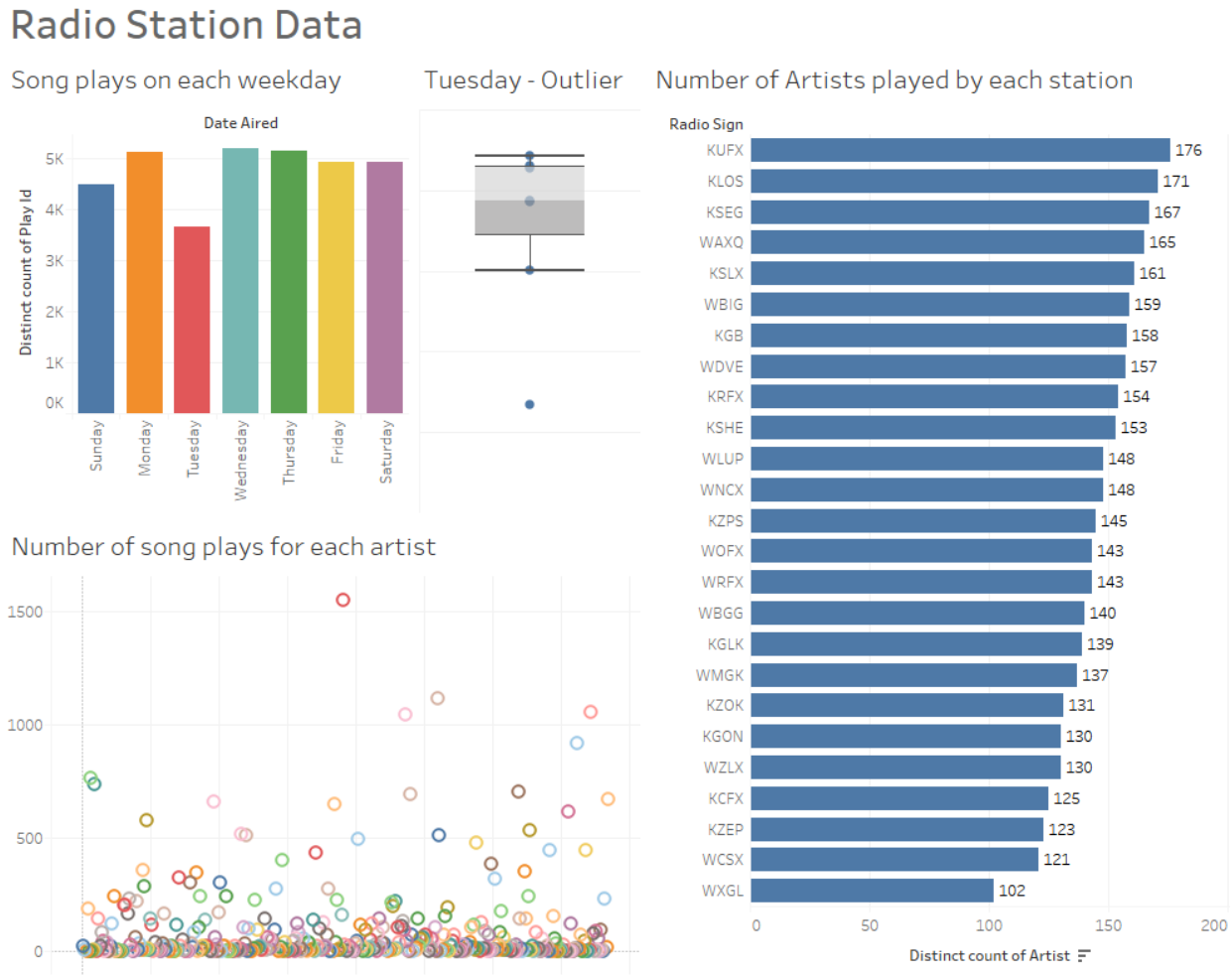
CREATE DATASET FOR USE IN EXTERNAL TOOLS

- I wish to look at the distribution of number of plays for songs.
- Saved `song`, `artist`, and `count(play_id)` as three columns in a new table.
- This column is exported into an xls file using the export data wizard.

ANALYSES AND VISUALIZATION USING EXTERNAL TOOLS

TABLEAU

- The data is directly imported into Tableau using the in-built connection method in tableau.
- I use the `'rock'` table for building visualization.
- Below is a snapshot of the dashboard built in tableau.



- Song plays on each weekday and Tuesday – outlier
  - The first bar graph shows the number of song plays over the entire day.
  - It is visually observable that the number of plays is low on Tuesdays.
  - To confirm this, the boxplot is drawn. It confirms that the outlier observed in the boxplot corresponds to Tuesday.
- Number of son plays for each artist
  - The jitter plot shows the number of song plays for each artist over the week.
  - It is an interactive way of looking for the top played artist.
  - The bottom part of the chart gives an estimate of how many plays artists have.
- Number of artists played by each station
  - It is seen that 'KUFX' owns rights to the largest number of artists a 176, while 'WXGL' has the least artists at 102.

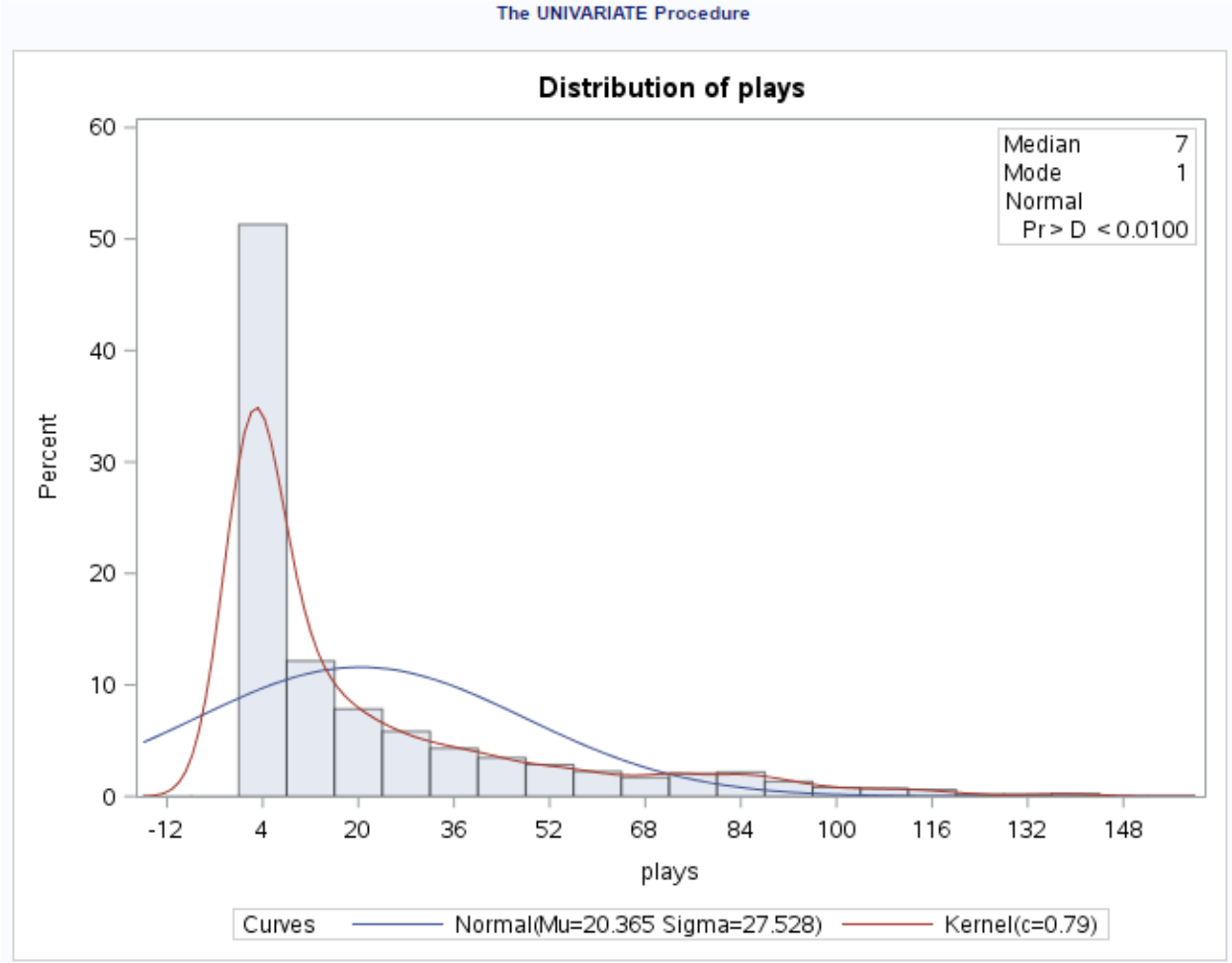
SAS

- An .xls file which is exported from SQL is imported into SAS using `proc import`.
- A `means` procedure has the following output –

The MEANS Procedure

Analysis Variable : plays plays						
Mean	Median	Mode	Lower Quartile	Upper Quartile	Skewness	Maximum
20.3649059	7.0000000	1.0000000	2.0000000	28.0000000	1.8066254	142.0000000

  - As can be seen, the data is heavily skewed, with mean at ~20 plays and mode at 1 play.
  - The skewness measure also says the same.
  - This is expected behavior – most songs are played only once a week. Some very popular songs get played more than 100 times in the same week.
  - The maximum played song was played 142 times over the week, or an average of 20 times every day. The song is 'Dream on' by Aerosmith. [From my personal knowledge of rock music, this song deserves being played 20 times a day every day in a week]
- The distribution of number of times a song is played is as shown below –



  - The red line is a kernel curve, estimating the distribution of the number of plays.
  - The blue line shows how a normal distribution with the estimated mean and standard deviation would look like.

DIFFICULTIES FACED WHILE HANDLING DATA

- I tried to export the same `'rock'` table in SAS, and perform analyses on aggregated data using SAS directly.
- However, I could not perform the analyses with the same ease as using aggregate data.
- So I created a new table using SQL and exported just the data needed for SAS analysis.