Bethel Tech
Christmas Challenge

Hans-Jörg Stark

December 2021

Task

Two questions to answer

1. ***What can you tell me about week position?***

2. ***What can you tell me about weeks on the chart?***

General Information on the data

- The dataset contains information on Christmas-songs on the chart from the 1950ies until the 2010s
- A unique entry is the combination of song-title and performer
- Individual songs can appear more than once because they were performed by different performers
- Individual performers can appear more than once because they can perform different songs in different seasons

General Information on the data

- 387 entries / records
- 13 attributes
- 70 unique songs
- 69 unique performers
- 78 unique combinations of performer & song

```
RangeIndex: 387 entries, 0 to 386
Data columns (total 13 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   url                     387 non-null    object
 1   weekid                  387 non-null    object
 2   week_position           387 non-null    int64
 3   song                    387 non-null    object
 4   performer               387 non-null    object
 5   songid                  387 non-null    object
 6   instance                387 non-null    int64
 7   previous_week_position  279 non-null    float64
 8   peak_position           387 non-null    int64
 9   weeks_on_chart          387 non-null    int64
 10  year                    387 non-null    int64
 11  month                   387 non-null    int64
 12  day                     387 non-null    int64
```

General Information on the data

- Some data-wrangling was done in MS Excel:
  I added a column that indicates whether the performer is an individual or a group
- Thereafter, I added some more variables to the dataset that were derived from the information given, like:
  - first position of a song before or after Christmas
  - Number of days to Christmas of a record
  - Does the title contain “Christmas”
  - etc. (cf. Jupyter Notebook)

General Information on the data

Data Problem I:

The peak-position does not match with the minimal week-position for 3 songs:

B	C	D	E	F	G	H	I	J	K	L	M	N
weekid	week_positi	song	performer	performerGrou	songid	instance	previous_week_positi	peak_position	weeks_on_cha	ye	mon	d
12/24/2005	97	BELIEVE	Brooks & Dunn	1	BelieveBrooks & Dunn	1		60	20	2005	12	24
12/31/2005	95	BELIEVE	Brooks & Dunn	1	BelieveBrooks & Dunn	1	97	60	20	2005	12	31
01.07.06	91	BELIEVE	Brooks & Dunn	1	BelieveBrooks & Dunn	1	95	60	20	2006	1	7
1/21/2006	90	BELIEVE	Brooks & Dunn	1	BelieveBrooks & Dunn	2		60	20	2006	1	21
1/28/2006	86	BELIEVE	Brooks & Dunn	1	BelieveBrooks & Dunn	2	90	60	20	2006	1	28
12/13/1980	75	SAME OLD LANG SYNE	Dan Fogelberg	0	Same Old Lang SyneDan Fogelberg	1		9	18	1980	12	13
12/20/1980	59	SAME OLD LANG SYNE	Dan Fogelberg	0	Same Old Lang SyneDan Fogelberg	1	75	9	18	1980	12	20
12/27/1980	37	SAME OLD LANG SYNE	Dan Fogelberg	0	Same Old Lang SyneDan Fogelberg	1	59	9	18	1980	12	27
01.03.81	37	SAME OLD LANG SYNE	Dan Fogelberg	0	Same Old Lang SyneDan Fogelberg	1	37	9	18	1981	1	3
01.10.81	31	SAME OLD LANG SYNE	Dan Fogelberg	0	Same Old Lang SyneDan Fogelberg	1	37	9	18	1981	1	10
1/17/1981	26	SAME OLD LANG SYNE	Dan Fogelberg	0	Same Old Lang SyneDan Fogelberg	1	31	9	18	1981	1	17
1/24/1981	19	SAME OLD LANG SYNE	Dan Fogelberg	0	Same Old Lang SyneDan Fogelberg	1	26	9	18	1981	1	24
1/31/1981	14	SAME OLD LANG SYNE	Dan Fogelberg	0	Same Old Lang SyneDan Fogelberg	1	19	9	18	1981	1	31
11.05.05	92	BETTER DAYS	Goo Goo Dolls	1	Better DaysGoo Goo Dolls	1	83	36	20	2005	11	5
11.12.05	77	BETTER DAYS	Goo Goo Dolls	1	Better DaysGoo Goo Dolls	1	92	36	20	2005	11	12
11/19/2005	71	BETTER DAYS	Goo Goo Dolls	1	Better DaysGoo Goo Dolls	1	77	36	20	2005	11	19
11/26/2005	67	BETTER DAYS	Goo Goo Dolls	1	Better DaysGoo Goo Dolls	1	71	36	20	2005	11	26
12.03.05	68	BETTER DAYS	Goo Goo Dolls	1	Better DaysGoo Goo Dolls	1	67	36	20	2005	12	3
12.10.05	65	BETTER DAYS	Goo Goo Dolls	1	Better DaysGoo Goo Dolls	1	68	36	20	2005	12	10
12/17/2005	57	BETTER DAYS	Goo Goo Dolls	1	Better DaysGoo Goo Dolls	1	65	36	20	2005	12	17
12/24/2005	59	BETTER DAYS	Goo Goo Dolls	1	Better DaysGoo Goo Dolls	1	57	36	20	2005	12	24
12/31/2005	57	BETTER DAYS	Goo Goo Dolls	1	Better DaysGoo Goo Dolls	1	59	36	20	2005	12	31
01.07.06	48	BETTER DAYS	Goo Goo Dolls	1	Better DaysGoo Goo Dolls	1	57	36	20	2006	1	7
1/14/2006	50	BETTER DAYS	Goo Goo Dolls	1	Better DaysGoo Goo Dolls	1	48	36	20	2006	1	14
1/21/2006	53	BETTER DAYS	Goo Goo Dolls	1	Better DaysGoo Goo Dolls	1	50	36	20	2006	1	21
1/28/2006	59	BETTER DAYS	Goo Goo Dolls	1	Better DaysGoo Goo Dolls	1	53	36	20	2006	1	28

Due to lack of knowledge on the data and their origin no action was taken

General Information on the data

Data Problem II:

There is a problem that some data is only available on song-ID level. For instance, the peak-position is the same for one song no matter whether it was more than once or just once on the charts (*instance* =1 or *instance* > 1).
The same is true of the *number of weeks* on the chart:
It is an overall sum, not a sum of the *number of weeks on chart for one season*. This makes it difficult to compare the data is some ways.

This is illustrated on the next slide.

The number of weeks on the chart per season for every song were computed and used this information to test whether a machine learning approach would be able to predict the time of presence of a song on the chart (at the very end of this presentation and the Jupyter Notebook).

General Information on the data

	weeksOnChartPerSeason	weeks_on_chart	instance	songid	peak_position
0	1	2	2	A Great Big SledThe Killers Featuring Toni Hal...	54
1	1	2	1	A Great Big SledThe Killers Featuring Toni Hal...	54
2	1	1	1	A Holly Jolly ChristmasBurl Ives	46
3	5	19	6	All I Want For Christmas Is YouMariah Carey	11
4	4	19	5	All I Want For Christmas Is YouMariah Carey	11
5	3	19	4	All I Want For Christmas Is YouMariah Carey	11
6	3	19	3	All I Want For Christmas Is YouMariah Carey	11
7	3	19	2	All I Want For Christmas Is YouMariah Carey	11
8	1	19	1	All I Want For Christmas Is YouMariah Carey	11
9	1	1	1	All I Want For Christmas Is YouMichael Buble	99
10	11	11	1	AmenThe Impressions	7
11	5	5	1	Auld Lang SyneKenny G	7
12	6	5	1	Baby's First ChristmasConnie Francis	26
13	2	2	1	Baby, It's Cold OutsideGlee Cast	57
14	3	20	1	BelieveBrooks & Dunn	60

What can you tell me about week position?

- The average initial position of a song is 75.3
- The overall average position of a song is 57.2

→ This means that on average the initial position of a song is worse or higher than the overall average position of all songs during their presence in the charts. In other words, a song on average improves its position.

What can you tell me about week position?

- The median initial position of a song is 82.
- The overall median position of a song is 58.

→ This confirms the findings based on average analysis: The initial position of a song is worse than the overall weekly position of all songs at all times.

What can you tell me about week position?

General statistical information on **week position**:

- Count: 387.000000
- Mean: 57.204134
- Stand Dev: 25.398527
- Min: 7.000000 = best position
- 25%: 38.500000
- 50%: 58.000000 = median position
- 75%: 78.000000
- Max: 100.000000 = worst position

What can you tell me about week position?

Which are the 10 most successful songs?

Assumption: success = largest interval from worst to best week-position

- **<u>Findings</u>**:
  The 10 most successful songs…
  - were all over the time of investigation, with a slight bend towards the earlier being more successful
  - had their initial position in 90% before Christmas
  - were no worse with their best position than below 30
  - were in 70% performed by individuals, not groups.

- The top 5 songs did not contain “Christmas” in their title

	song	performer	timestamp	MinWeekPosition	MaxWeekPosition	weekPositionRange	AverageWeekPosition
0	AMEN	The Impressions	1964-11-21	7	96	89	31.818182
1	MISTLETOE	Justin Bieber	2011-12-3	11	96	85	50.900000
2	AULD LANG SYNE	Kenny G	1999-12-25	7	89	82	44.800000
3	JINGLE BELL ROCK	Bobby Rydell/Chubby Checker	1962-12-15	21	97	76	49.625000
4	THIS ONE'S FOR THE CHILDREN	New Kids On The Block	1989-11-11	7	82	75	26.250000
5	ROCKIN' AROUND THE CHRISTMAS TREE	Brenda Lee	1962-12-15	14	89	75	49.210526
6	WHITE CHRISTMAS	Bing Crosby	1958-12-20	12	86	74	44.357143
7	ALL I WANT FOR CHRISTMAS IS YOU	Mariah Carey	2000-1-8	11	83	72	28.947368
8	JINGLE BELL ROCK	Bobby Helms	1962-12-8	29	99	70	53.650000
9	IF WE MAKE IT THROUGH DECEMBER	Merle Haggard	1973-11-24	28	97	69	59.600000

What can you tell me about week position?

Which are the 10 most successful songs?

Assumption: success = best week-position ever during their presence in the charts

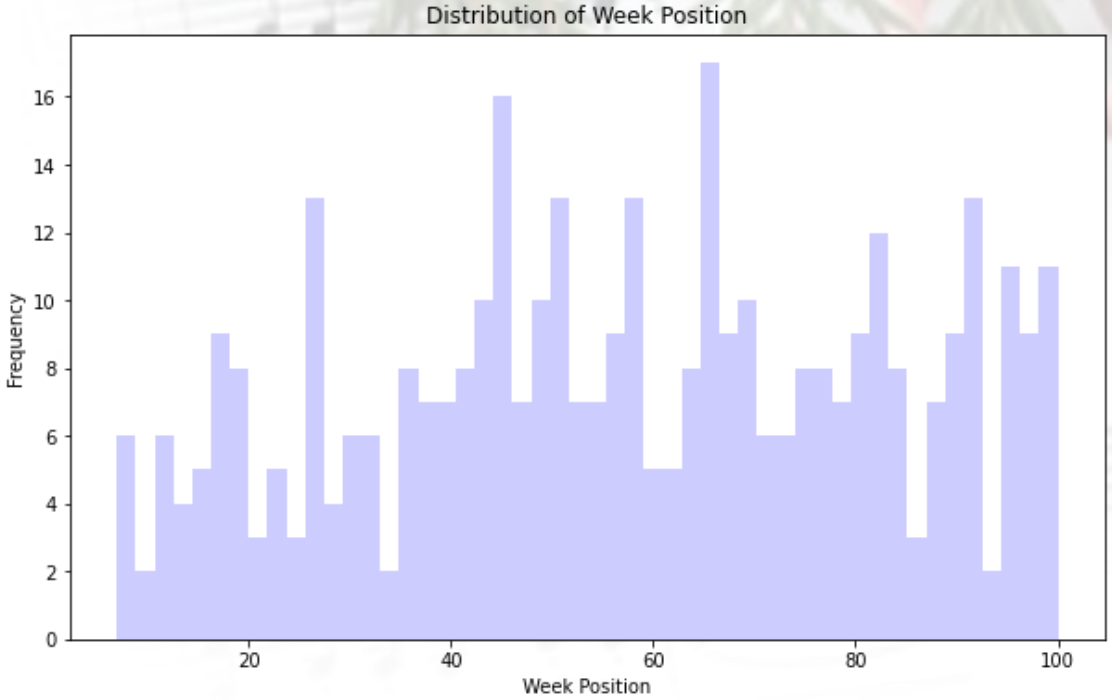
- **Findings**:

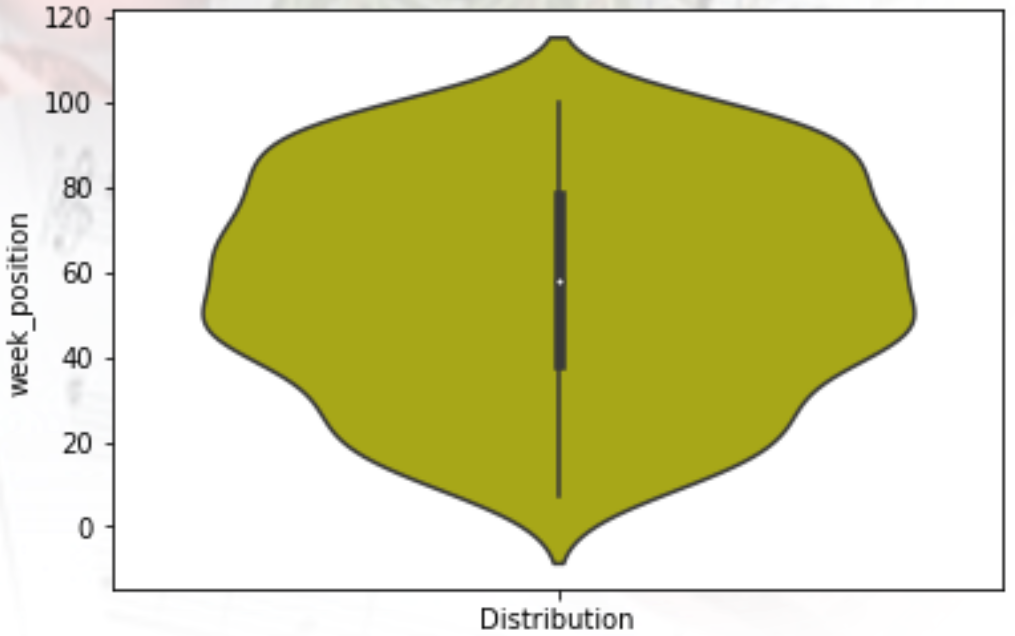
  The 10 most successful songs…
  - were all over the time of investigation
  - had their initial position in 90% before or at Christmas
  - were at least during 5 weeks on the charts, in 70% at least 10 weeks
  - were performed in 80% by individuals, not groups.
  - Contained in 40% “Christmas” in their title

song	performer	timestamp	PeakPosition	eeksOnChart	MinWeekPosition	MaxWeekPosition	weekPositionRange	AverageWeekPosition
AMEN	The Impressions	1964-11-21	7	11	7	96	89	31.818182
AULD LANG SYNE	Kenny G	1999-12-25	7	5	7	89	82	44.800000
THIS ONE'S FOR THE CHILDREN	New Kids On The Block	1989-11-11	7	16	7	82	75	26.250000
SAME OLD LANG SYNE	Dan Fogelberg	1980-12-13	9	18	14	75	61	37.250000
ALL I WANT FOR CHRISTMAS IS YOU	Mariah Carey	2000-1-8	11	19	11	83	72	28.947368
MISTLETOE	Justin Bieber	2011-12-3	11	10	11	96	85	50.900000
WHITE CHRISTMAS	Bing Crosby	1958-12-20	12	13	12	86	74	44.357143
DO THEY KNOW IT'S CHRISTMAS?	Band-Aid	1984-12-22	13	9	13	65	52	32.500000
ROCKIN' AROUND THE CHRISTMAS TREE	Brenda Lee	1962-12-15	14	18	14	89	75	49.210526
PRETTY PAPER	Roy Orbison	1963-12-14	15	7	15	60	45	30.285714

What can you tell me about week position?

Graphical statistical information on **week position**:





What can you tell me about week position?

- There are 279 records out of 387 (~72%) that indicate a change in chart's position.
- The remaining were single presences

What can you tell me about week position?

Is the change in ranking better before or after Christmas?

Change in rankings are in months November, December and January.

- 81 entries with better rankings before Christmas (40.5%)
- 119 entries with better rankings after Christmas (59.5%)

What can you tell me about week position?

Null Hypothesis H0 : The average week-position is independent of whether the week-position is before or after Christmas. i.e. the two group means of the groups – before and after Christmas – are the same.

→Run a t-test: p-value = 3.648e-09

H0 is being rejected. **The average week-position before Christmas differs significantly from the average week-position after Christmas!**

Mean Week-Position before Xmas: **65.6** | Mean Week-Position after Xmas: **50.6**

What can you tell me about week position?

Counting the numbers of datasets that have “Christmas” in their title:

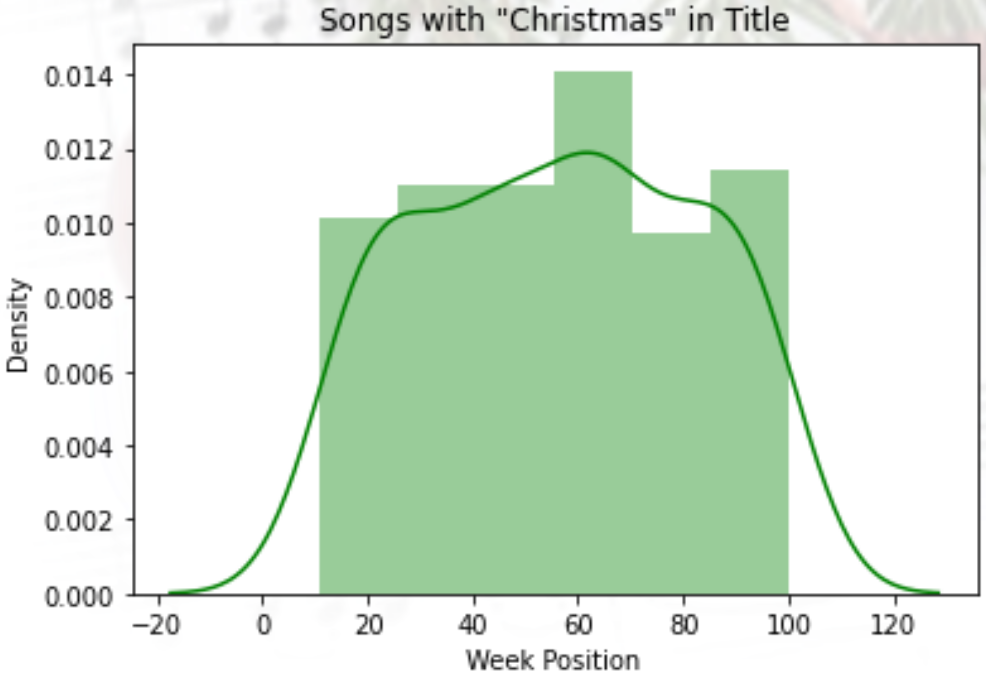
- 234 (60.5%) without “Christmas”
- 153 (39.5%) with “Christmas”

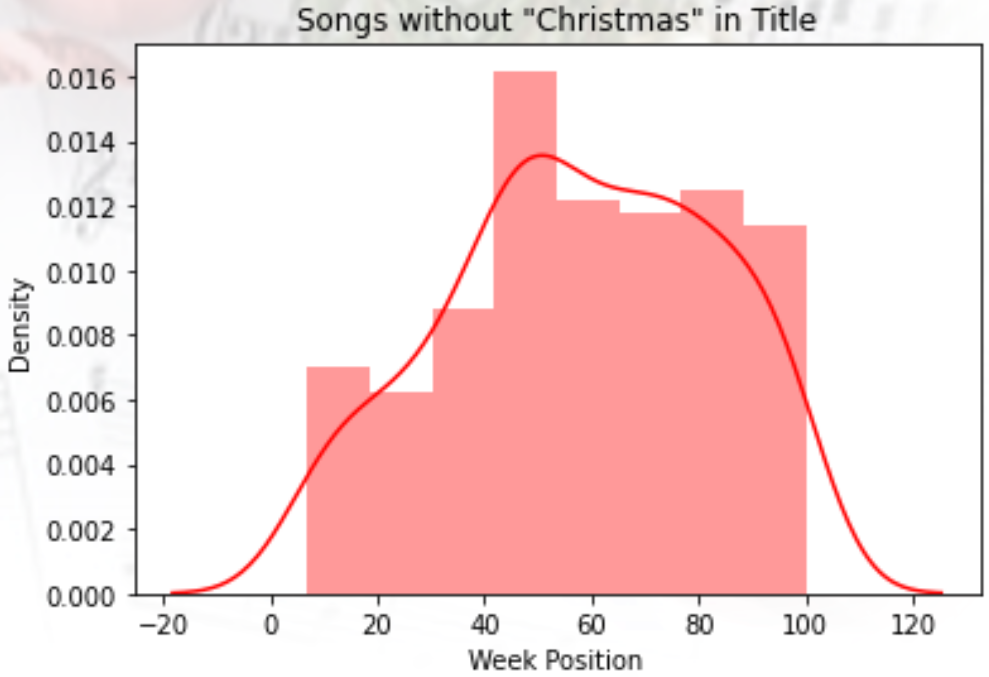
Counting the numbers of unique songs that have “Christmas” in their title:

- 44 (56.4%) without “Christmas”
- 34 (43.6%) with “Christmas”

What can you tell me about week position?

Some graphic information on the split:





What can you tell me about week position?

Null Hypothesis H0 : The average week-position is independent of whether the song contains “Christmas” in its title.

→Run a t-test: p-value = 0.570

H0 is being accepted. **The average week-position of a song is not dependant on whether the songtitle contains "Christmas"**

What can you tell me about week position?

Null Hypothesis H0 : Whether a song is performed by an individual or a group does not have an influence on the week-position in the chart.

→Run a t-test: p-value = 0.961

H0 is being accepted. **There is no indication that the performer has a signifcant influence on the week-position of a song in the chart.**

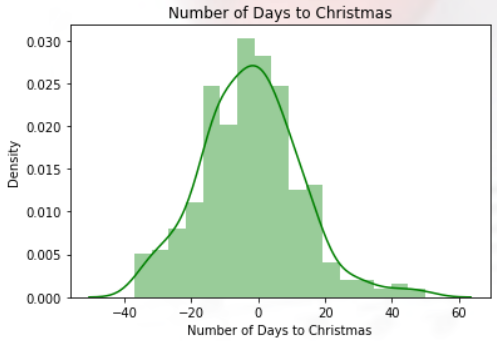
What can you tell me about week position?

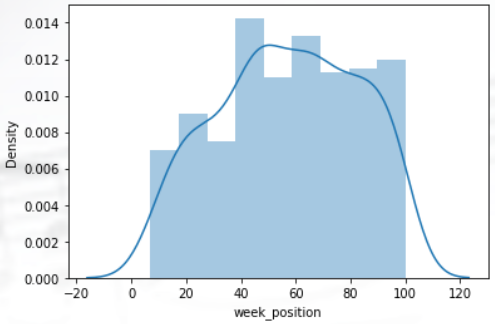
Influence of Time to Christmas on Week Position: Does the closeness (timewise) to Christmas have an influence on week-position?

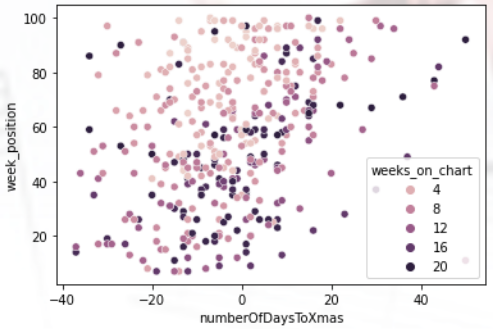
Null Hypothesis H0: The time-distance to Christmas has no significant influence on the week-position of a song in the charts.

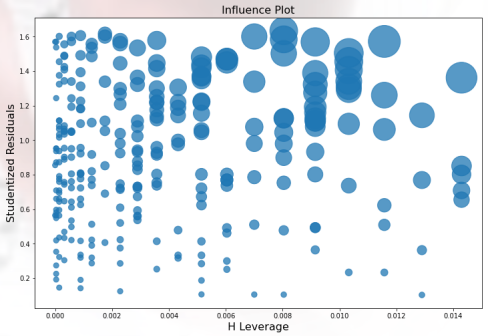
→ With a p-value = 0.93 H0 is accepted!

(Running a linear regression)









What can you tell me about week position?

Test if the good or bad week-position differs significantly during the decades. For this test the week-position needed to be recoded because there were too few data. The recoding was: 0 for a position below median (<58 as good) and 1 (>58 as bad) above.

Run an Independent Chi-Square Test

Null Hypothesis H0: The decade has no significant influence on the quality (good/bad) of week-position of a song.

→**With a p-value = 0.00507 H0 is rejected!**

In other words: The decade has a significant influence on the quality of week position of a song!

weekPosCat50	0	1	All
decade			
1950	7	14	21
1960	83	61	144
1970	19	25	44
1980	17	18	35
1990	8	16	24
2000	13	37	50
2010	48	21	69
All	195	192	387

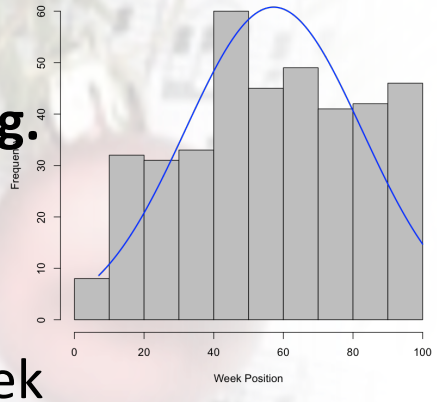
What can you tell me about week position?

Test if the week-position differs significantly during the decades.

Run a Repeated ANOVA Test in R:

Null Hypothesis H0: The decade of a songs appearance in the chart has no significant influence on the week-position of a song.

→**With a p-value = 0.011 H0 is rejected!**



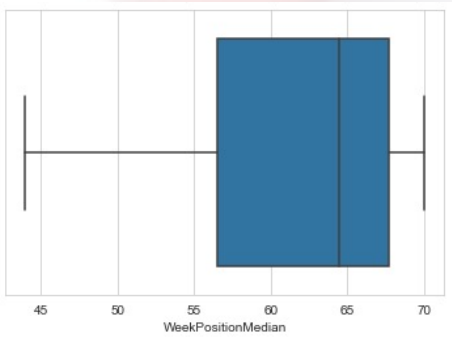
In other words: The decade has a significant influence on the week position of a song in the charts! This result confirms the finding and result of the Chi Square test conducted before.

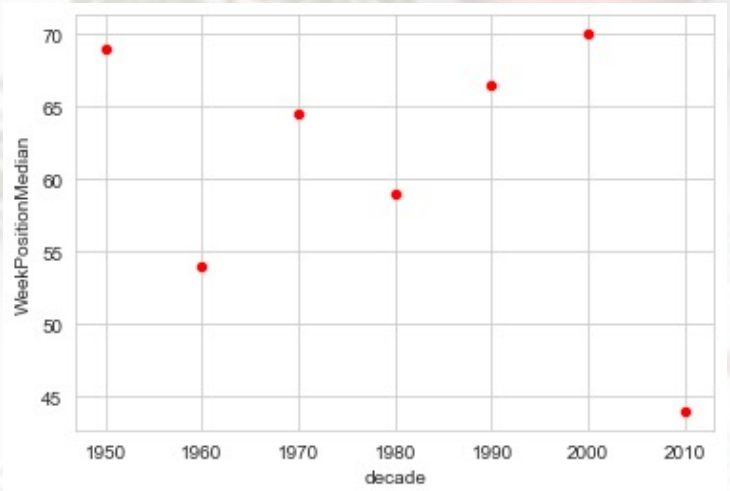
What can you tell me about week position?

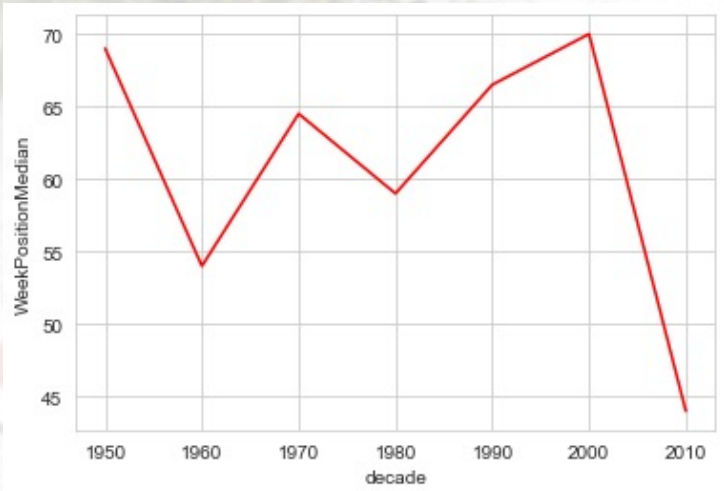
Crosstable and visual investigation of the median week position per decade.

Correlation = -0.337

WeekPositionMedian	44.0	54.0	59.0	64.5	66.5	69.0	70.0	All
decade								
1950.0	0	0	0	0	0	1	0	1
1960.0	0	1	0	0	0	0	0	1
1970.0	0	0	0	1	0	0	0	1
1980.0	0	0	1	0	0	0	0	1
1990.0	0	0	0	0	1	0	0	1
2000.0	0	0	0	0	0	0	1	1
2010.0	1	0	0	0	0	0	0	1
All	1	1	1	1	1	1	1	7







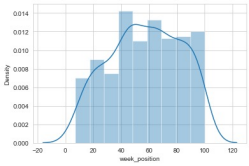
Interestingly the median drops significantly for the decade 2010!

What can you tell me about week position?

Test if the month of rank has a significant influence on the week-position

Run an ANOVA Test:

Null Hypothesis H0: The month has no significant influence on the week-position of a song in the charts.



→With a p-value = 0.000034 H0 is rejected!

The month has a significant influence on the week position of a song in the charts! A Welch's ANOVA test revealed that the mean difference between:

- **December and January**
- **November and January**

are significantly different, with the best average position in *January*, followed by *December* and *November*.

A	B	mean(A)	mean(B)	diff	se	T	df	pval	hedges
Dec	Jan	61.508929	49.162162	12.346766	2.627381	4.699267	304.571722	0.001000	0.496781
Dec	Nov	61.508929	72.266667	-10.757738	6.572620	-1.636750	15.842499	0.259827	-0.435145
Jan	Nov	49.162162	72.266667	-23.104505	6.701693	-3.447562	17.110140	0.008113	-0.929819

What can you tell me about week position?

	MinWeekPosition	MaxWeekPosition	AverageWeekPosition	weekPosRange	songid
0	7	96	31.818182	89	AmenThe Impressions
1	11	96	50.900000	85	MistletoeJustin Bieber
2	7	89	44.800000	82	Auld Lang SyneKenny G
3	21	97	49.625000	76	Jingle Bell RockBobby Rydell/Chubby Checker
4	7	82	26.250000	75	This One's For The ChildrenNew Kids On The Block
...	...	...	...	...	...
73	92	92	92.000000	0	Do They Know It's Christmas?Glee Cast
74	95	95	95.000000	0	Child Of GodBobby Darin
75	97	97	97.000000	0	Blue ChristmasThe Browns Featuring Jim Edward ...
76	99	99	99.000000	0	All I Want For Christmas Is YouMichael Buble
77	46	46	46.000000	0	A Holly Jolly ChristmasBurl Ives

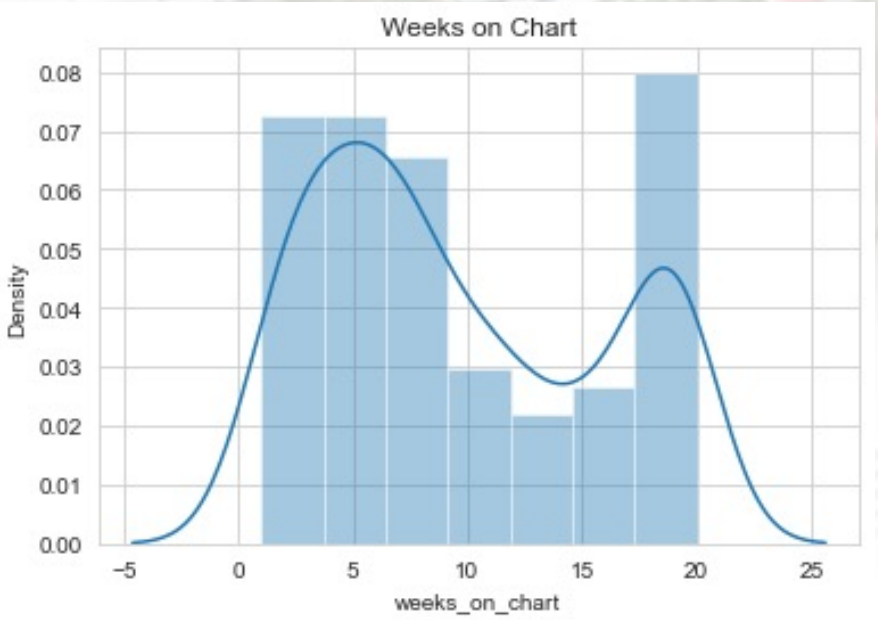
What can you tell me about weeks on the chart?

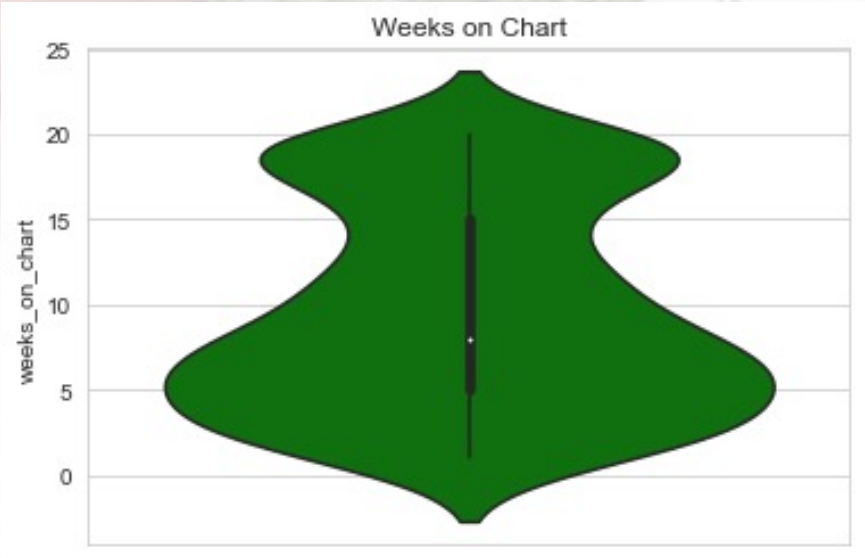
General statistical information on **Weeks on Chart (WoC) globally**:

- Count: 387.000000
- Mean: 9.645995
- Stand Dev: 6.142627
- Min: 1.000000 = shortest time of presence in chart
- 25%: 5.000000
- 50%: 8.000000 = median time of presence in chart
- 75%: 15.000000
- Max: 20.000000 = longest time of presence in chart

What can you tell me about weeks on the chart?

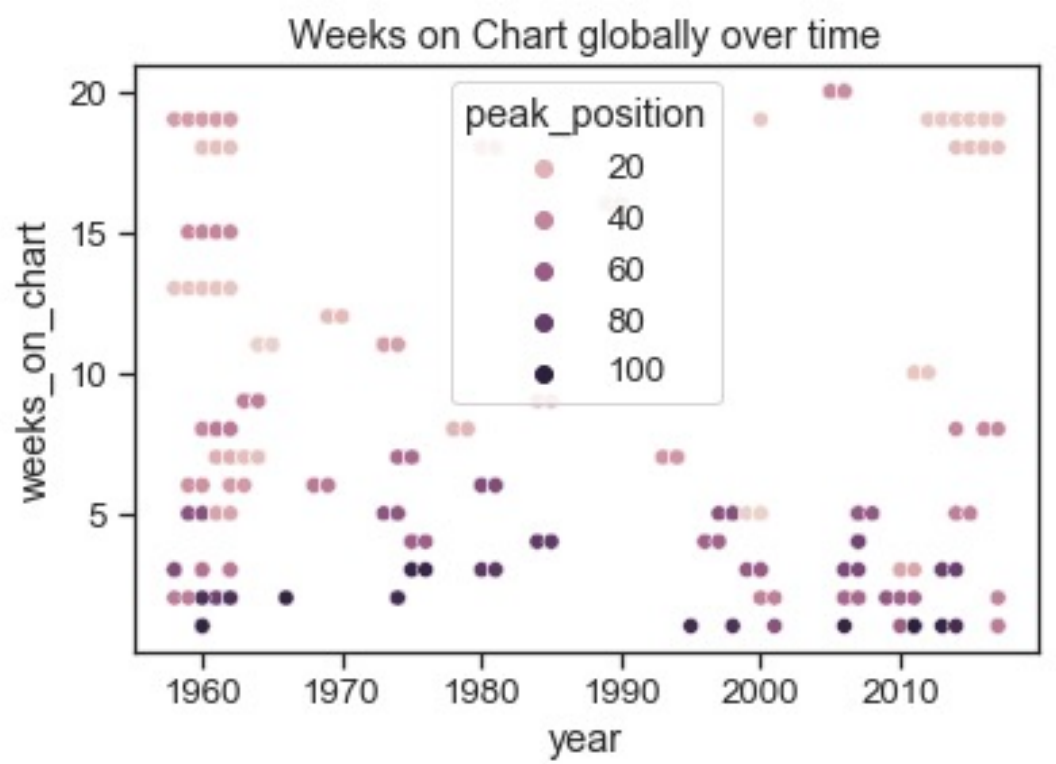
General statistical information on **Weeks on Chart (WoC) globally**:





What can you tell me about weeks on the chart?

General statistical information on **Weeks on Chart (WoC) globally**:

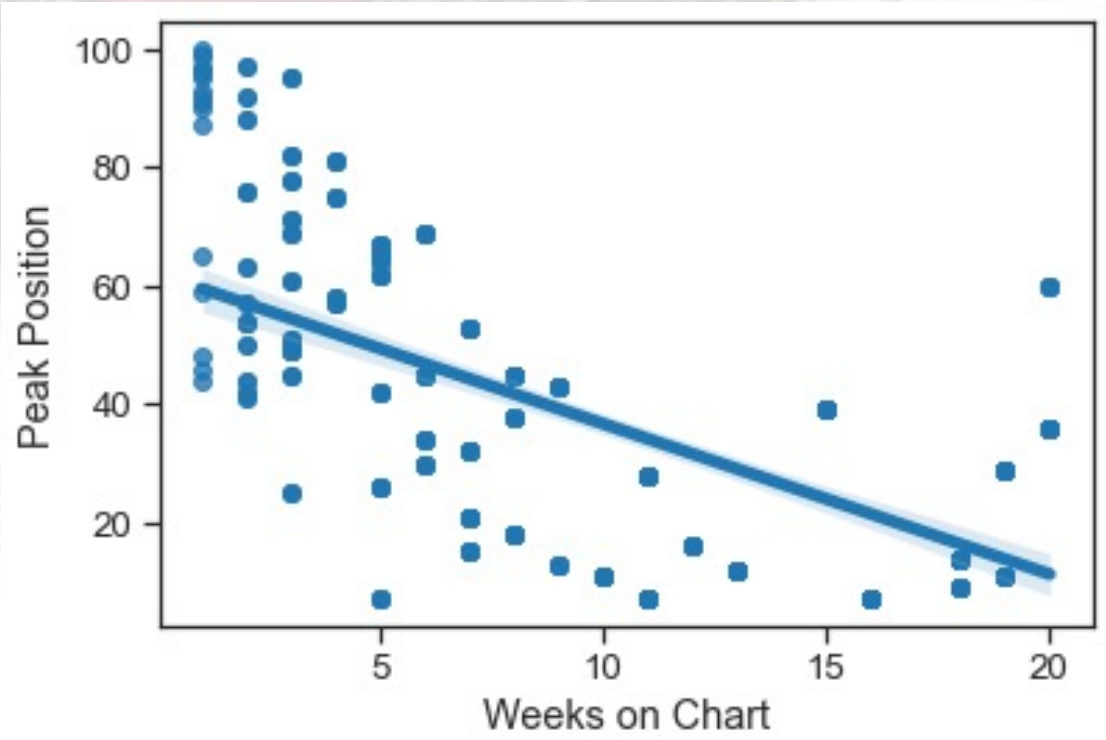


What can you tell me about weeks on the chart?

General statistical information on **Weeks on Chart (WoC) globally**:

The longer a song is on the chart, the better its position (i.e. the lower the number of the peak position)
→ negative correlation

which is to be expected



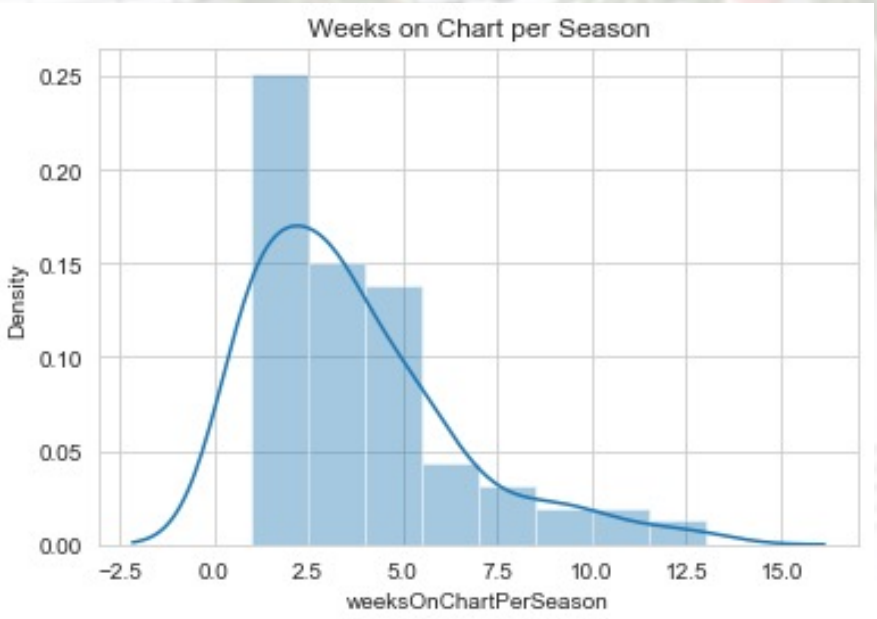
What can you tell me about weeks on the chart?

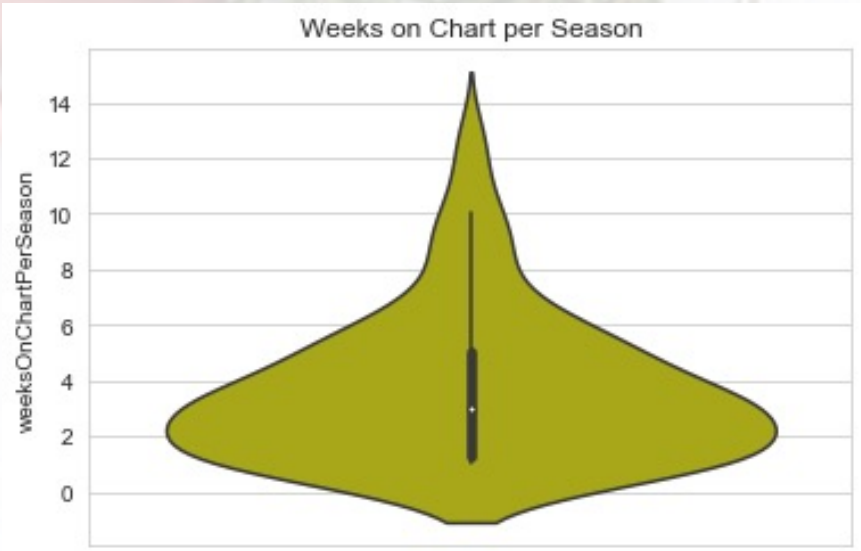
General statistical information on **Weeks on Chart (WoC) per Season**:

- Count: 106.000000
- Mean: 3.650943
- Stand Dev: 2.665628
- Min: 1.000000 = shortest time of presence in chart
- 25%: 1.250000
- 50%: 3.000000 = median time of presence in chart
- 75%: 5.000000
- Max: 13.000000 = longest time of presence in chart

What can you tell me about weeks on the chart?

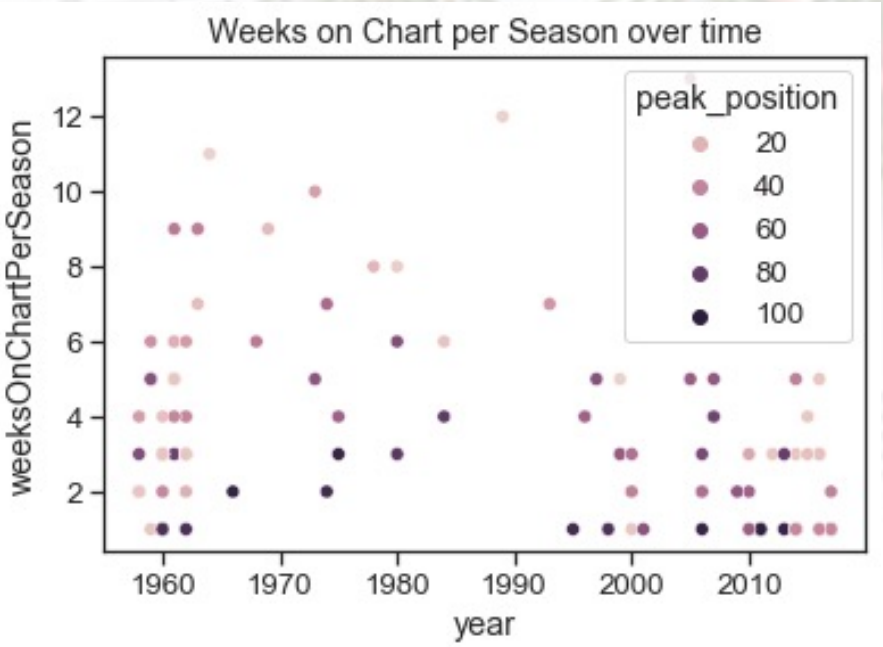
General statistical information on **Weeks on Chart (WoC) Per Season**:

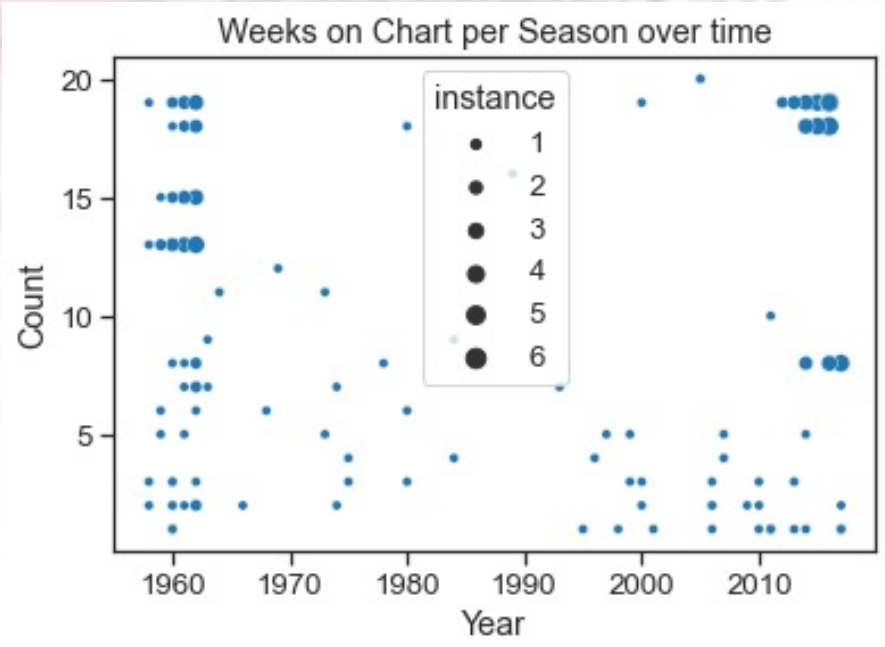




What can you tell me about weeks on the chart?

General statistical information on **Weeks on Chart (WoC) per Season**:



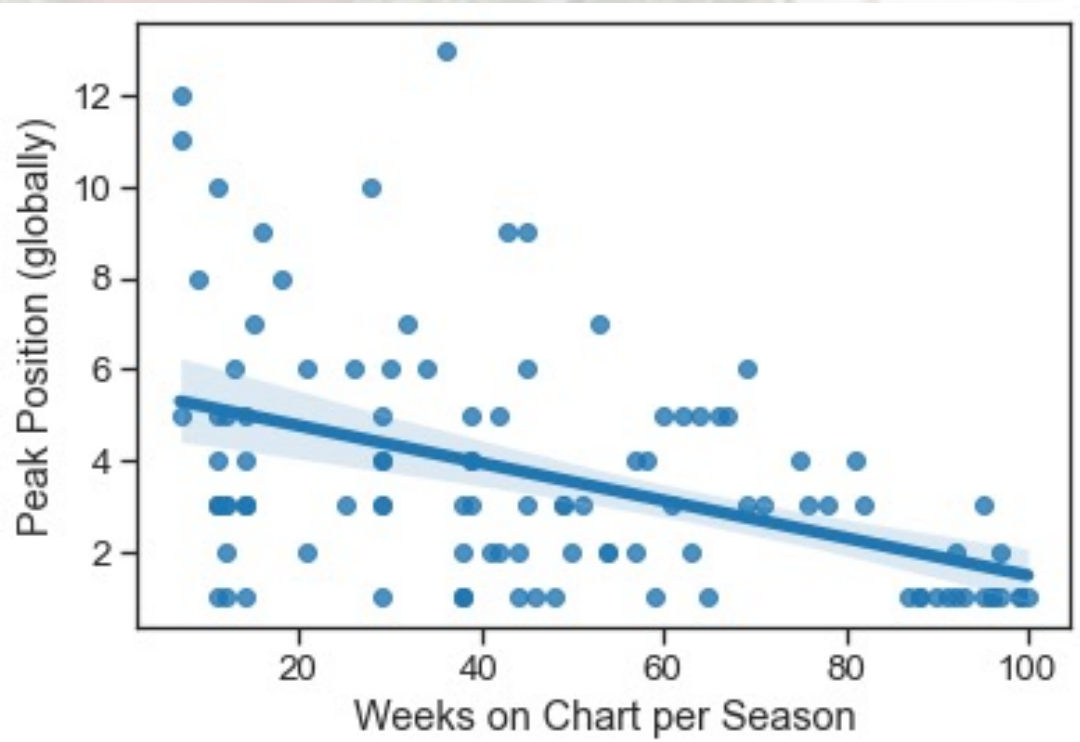


What can you tell me about weeks on the chart?

General statistical information on **Weeks on Chart (WoC) per Season**:

The longer a song is on the chart per season, the better its position (i.e. the lower the number of the peak position)
→ negative correlation

which is to be expected, although not as strong as globally



What can you tell me about weeks on the chart?

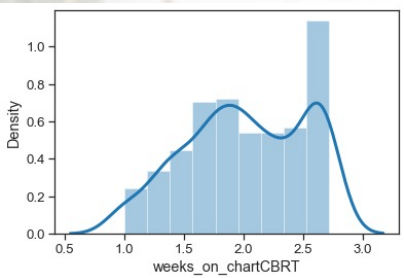
Null Hypothesis H0 : There is no signifcant difference in the means of the songs performed by a group or an individual.

.

→**With a p-value = 0.195 H0 is accepted!**

What can you tell me about weeks on the chart?

Null Hypothesis H0 : There is no siginifcant influence of decade of the Song position on the weeks of presence in the Chart.



→**With a p-value = 0.0243 H0 is rejected!**

According to Welch's ANOVA:

	A	B	mean(A)	mean(B)	diff	se	T		pval	hedges
0	1950	1960	2.031489	2.066530	-0.035041	0.115885	-0.302381	24.24289	0.900000	070307
1	1950	1970	2.031489	1.927808	0.103681	0.119317	0.868957	26.93131	0.900000	227718
2	1950	1980	2.031489	2.148121	-0.116632	0.132318	-0.881455	37.10802	0.900000	239910
3	1950	1990	2.031489	1.813400	0.218089	0.138072	1.579533	38.32290	0.672758	463695
4	1950	2000	2.031489	1.937693	0.093796	0.142300	0.659142	46.83797	0.900000	169531
5	1950	2010	2.031489	2.012137	0.019352	0.135717	0.142594	42.41208	0.900000	035234
6	1960	1970	2.066530	1.927808	0.138723	0.057344	2.419148	99.85517	0.201557	415027
7	1960	1980	2.066530	2.148121	-0.081591	0.080992	-1.007394	51.04690	0.900000	189044
8	1960	1990	2.066530	1.813400	0.253131	0.090087	2.809861	31.88158	0.104628	616713
9	1960	2000	2.066530	1.937693	0.128837	0.096441	1.335911	64.71711	0.811828	218429
10	1960	2010	2.066530	2.012137	0.054394	0.086433	0.629314	95.96496	0.900000	091813
11	1970	1980	1.927808	2.148121	-0.220314	0.085832	-2.566812	58.37682	0.155410	575682
12	1970	1990	1.927808	1.813400	0.114408	0.094461	1.211161	36.98691	0.879067	303838
13	1970	2000	1.927808	1.937693	-0.009886	0.100540	-0.098327	71.78078	0.900000	020159
14	1970	2010	1.927808	2.012137	-0.084329	0.090984	-0.926858	102.53105	0.900000	177603
15	1980	1990	2.148121	1.813400	0.334722	0.110428	3.031127	51.50256	0.054821	792707
16	1980	2000	2.148121	1.937693	0.210428	0.115671	1.819199	82.94996	0.533772	397298
17	1980	2010	2.148121	2.012137	0.135985	0.107469	1.265344	95.07461	0.853024	260648
18	1990	2000	1.813400	1.937693	-0.124294	0.122211	-1.017040	65.97726	0.900000	249919
19	1990	2010	1.813400	2.012137	-0.198737	0.114479	-1.736018	65.40583	0.582532	408001
20	2000	2010	1.937693	2.012137	-0.074443	0.119544	-0.622729	107.66991	0.900000	114912

Interestingly there is according to the p-value no significant difference between the means of the decades. The closest to a significant difference is between the 80ies and 90ies.

decadeR	decade	weeks_on_chart	weeks_on_chartCBRT
1950ies	1950.0	9.857143	2.031489
1960ies	1960.0	9.895833	2.066530
1970ies	1970.0	7.659091	1.927808
1980ies	1980.0	11.057143	2.148121
1990ies	1990.0	6.833333	1.813400
2000ies	2000.0	9.640000	1.937693
2010ies	2010.0	10.594203	2.012137

What can you tell me about weeks on the chart?

Null Hypothesis H0 : There is no signifcant influence of peak-position of the Song on the weeks of presence in the Chart.

→**With a p-value = 9.42e-55 H0 is rejected!**

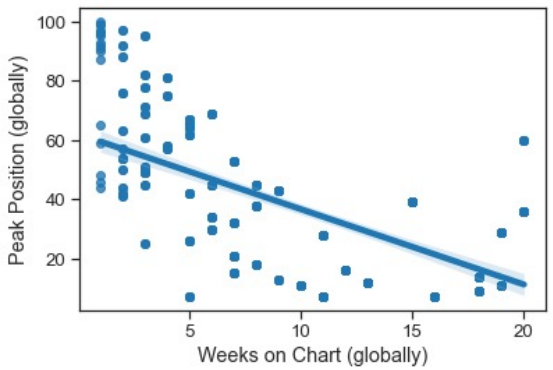
OLS Regression Results

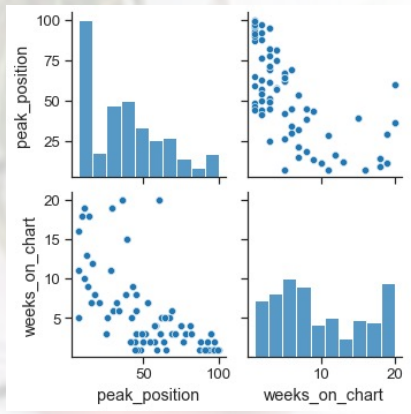
Dep. Variable:	weeks_on_chart	**R-squared (uncentered):**	0.467
Model:	OLS	**Adj. R-squared (uncentered):**	0.466
Method:	Least Squares	**F-statistic:**	338.7
Date:	Fri, 31 Dec 2021	**Prob (F-statistic):**	9.42e-55
Time:	18:20:18	**Log-Likelihood:**	-1370.1
No. Observations:	387	**AIC:**	2742.
Df Residuals:	386	**BIC:**	2746.
Df Model:	1		
Covariance Type:	nonrobust		

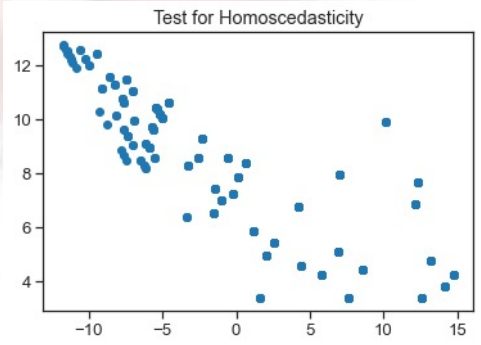
	coef	std err	t	P>|t|	[0.025	0.975]
peak_positionSQRT	1.2756	0.069	18.404	0.000	1.139	1.412

Omnibus:	518.565	**Durbin-Watson:**	0.377
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	27.467
Skew:	0.052	**Prob(JB):**	1.09e-06
Kurtosis:	1.699	**Cond. No.**	1.00

	songid	peak_position	weeks_on_chart
0	A Great Big SledThe Killers Featuring Toni Hal...	54	2
1	A Holly Jolly ChristmasBurl Ives	46	1
2	All I Want For Christmas Is YouMariah Carey	11	19
3	All I Want For Christmas Is YouMichael Buble	99	1
4	AmenThe Impressions	7	11







What can you tell me about weeks on the chart?

Null Hypothesis H0 : There is no signifcant influence of the month of first appearance on the chart of the Song on the weeks of presence in the Chart.

→**With a p-value = 0.003H0 is rejected!**

The p-values in the table to the right show:
The mean difference between
- **January and November and**
- **November and December**

are significantly different.

	A	B	mean(A)	mean(B)	diff	se	T	df	pval	hedges
0	1	11	9.155405	13.866667	-4.711261	1.277051	-3.689171	19.238816	0.004161	-0.994982
1	1	12	9.155405	9.687500	-0.532095	0.645030	-0.824914	324.378940	0.672241	-0.087205
2	11	12	13.866667	9.687500	4.179167	1.250446	3.342140	17.718434	0.009785	0.888538

Also: the average duration of presence on the chart is longest if the song first appeared in November, then in December and finally in January - which was to be expected. Although the difference between December and January is minimal.

What can you tell me about weeks on the chart?

Null Hypothesis H0 : There is no signifcant influence of of time-to-christmas of first appearance of the Song on the weeks of presence in the Chart.

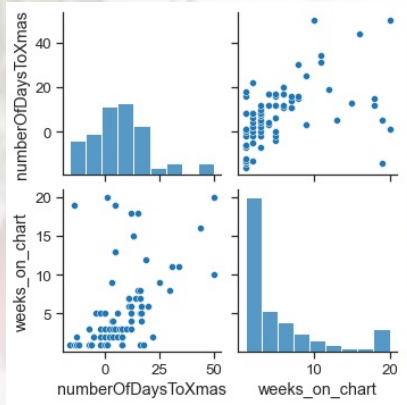
→**With a p-value = 6.09e-54H0 is rejected!**

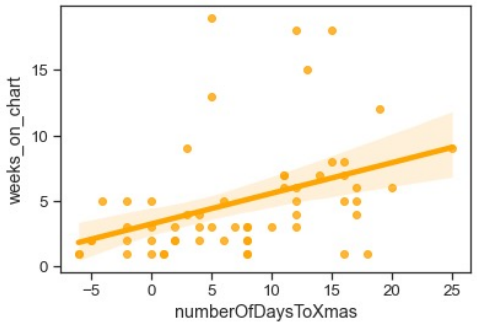
OLS Regression Results

Dep. Variable:	weeks_on_chartCBRT	**R-squared (uncentered):**	0.984
Model:	OLS	**Adj. R-squared (uncentered):**	0.984
Method:	Least Squares	**F-statistic:**	3612.
Date:	Fri, 31 Dec 2021	**Prob (F-statistic):**	6.09e-54
Time:	18:34:32	**Log-Likelihood:**	-24.538
No. Observations:	59	**AIC:**	51.08
Df Residuals:	58	**BIC:**	53.15
Df Model:	1		
Covariance Type:	nonrobust		

The number of days to (or from) Christmas explain **98.4%** (!) of the variability in number of weeks on chart (R-squared value).

The number of days to (or from) Christmas significantly influence the number of weeks on chart.





What can you tell me about weeks on the chart?

Null Hypothesis H0 : There is no significant influence of instances of a Song on the weeks of presence in the Chart.

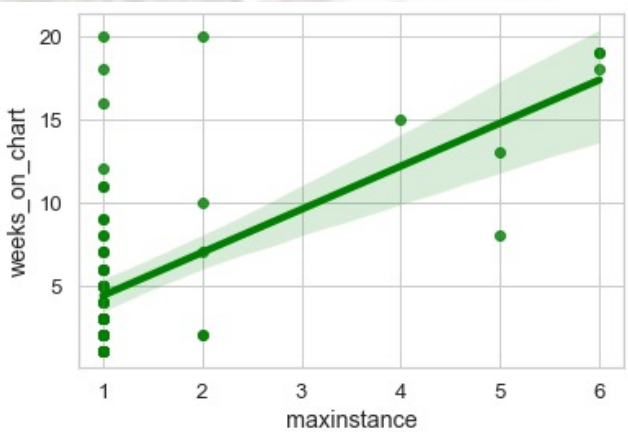
→**With a p-value = 7.10e-42 H0 is rejected!**

OLS Regression Results

Dep. Variable:	weeks_on_chartCBRT	**R-squared (uncentered):**	0.954
Model:	OLS	**Adj. R-squared (uncentered):**	0.953
Method:	Least Squares	**F-statistic:**	1250.
Date:	Fri, 31 Dec 2021	**Prob (F-statistic):**	7.10e-42
Time:	18:39:35	**Log-Likelihood:**	-16.264
No. Observations:	61	**AIC:**	34.53
Df Residuals:	60	**BIC:**	36.64
Df Model:	1		
Covariance Type:	nonrobust		

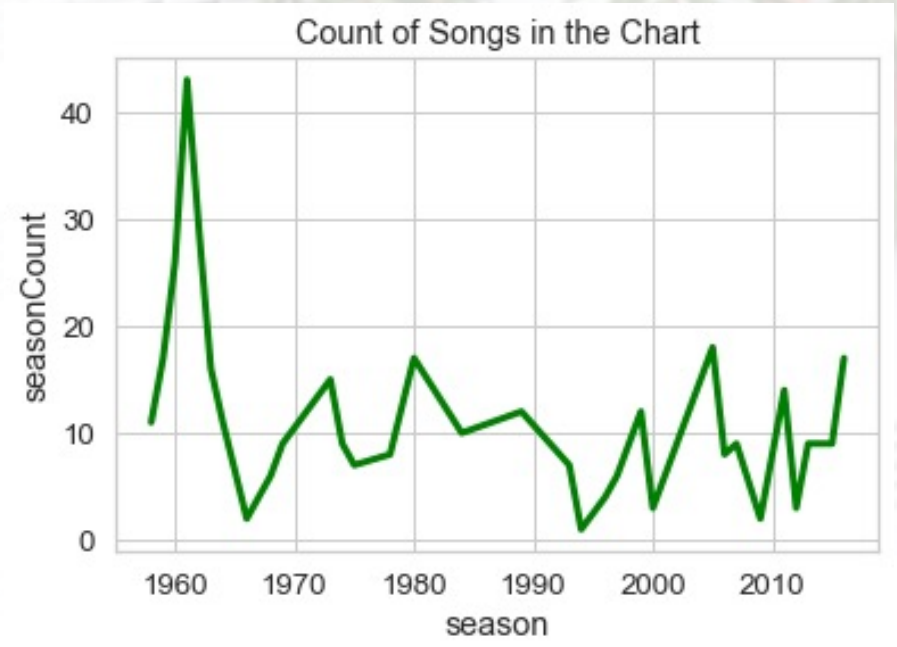
The (max) number of instances explains **95.4%** of the variability in number of weeks on chart (R-squared value).

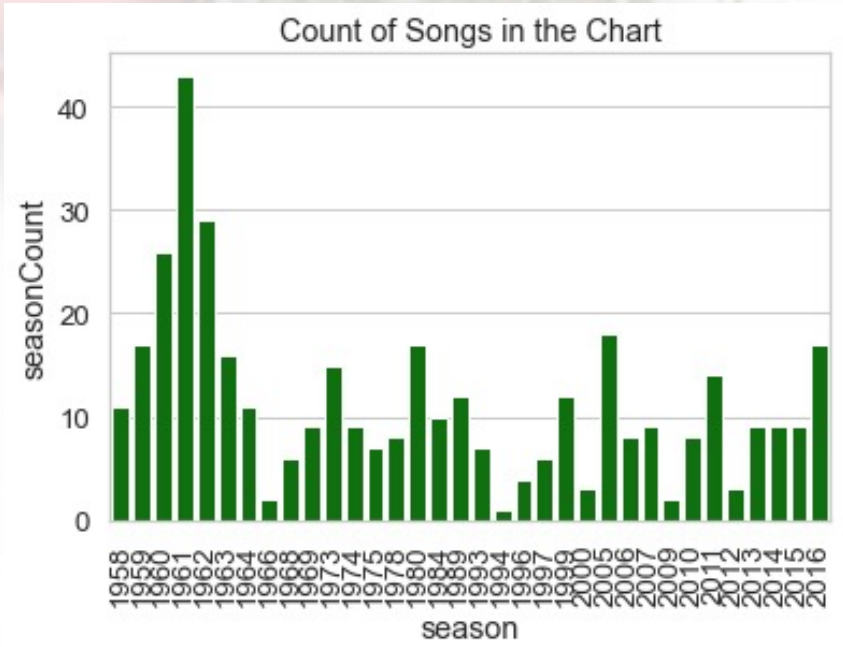
The (max) number of instances significantly influence the number of weeks on chart.



What can you tell me about weeks on the chart?

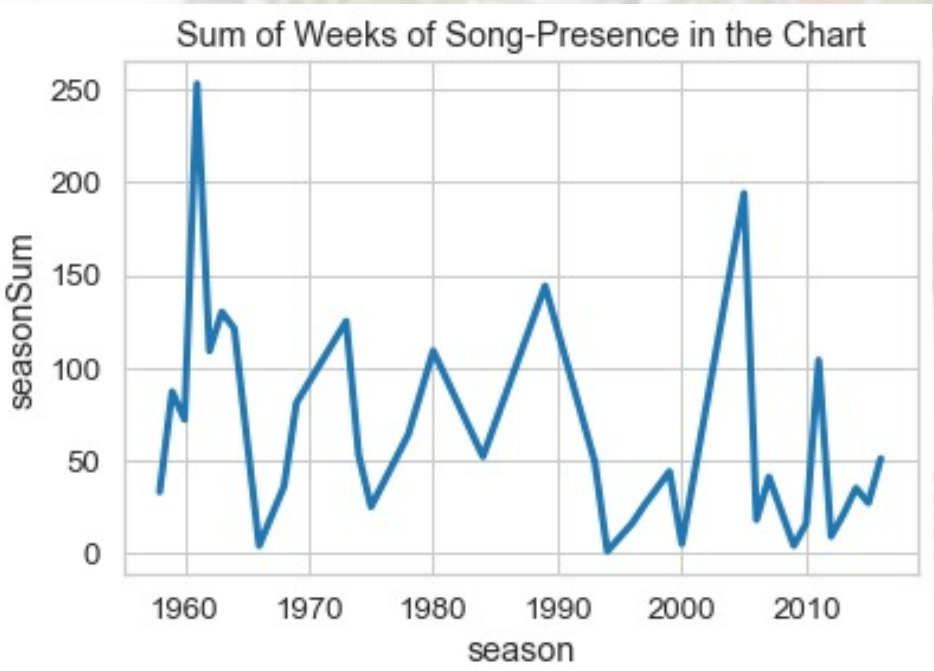
Some statistics on song counts per season

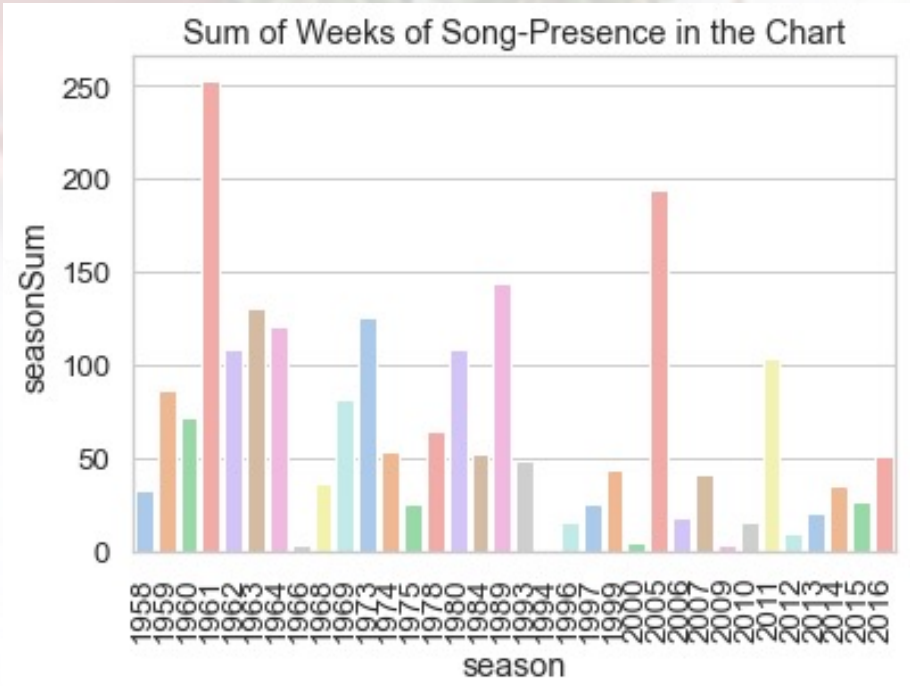




What can you tell me about weeks on the chart?

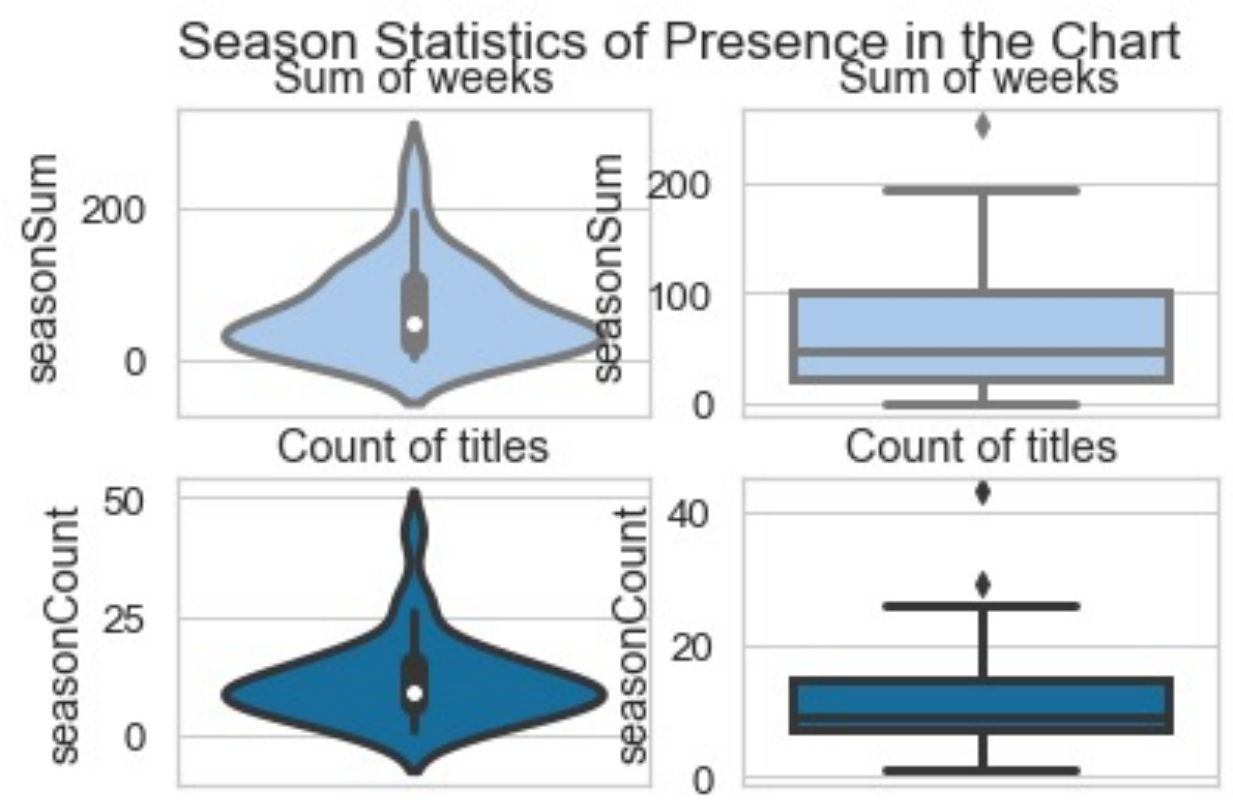
Some statistics on song presence





What can you tell me about weeks on the chart?

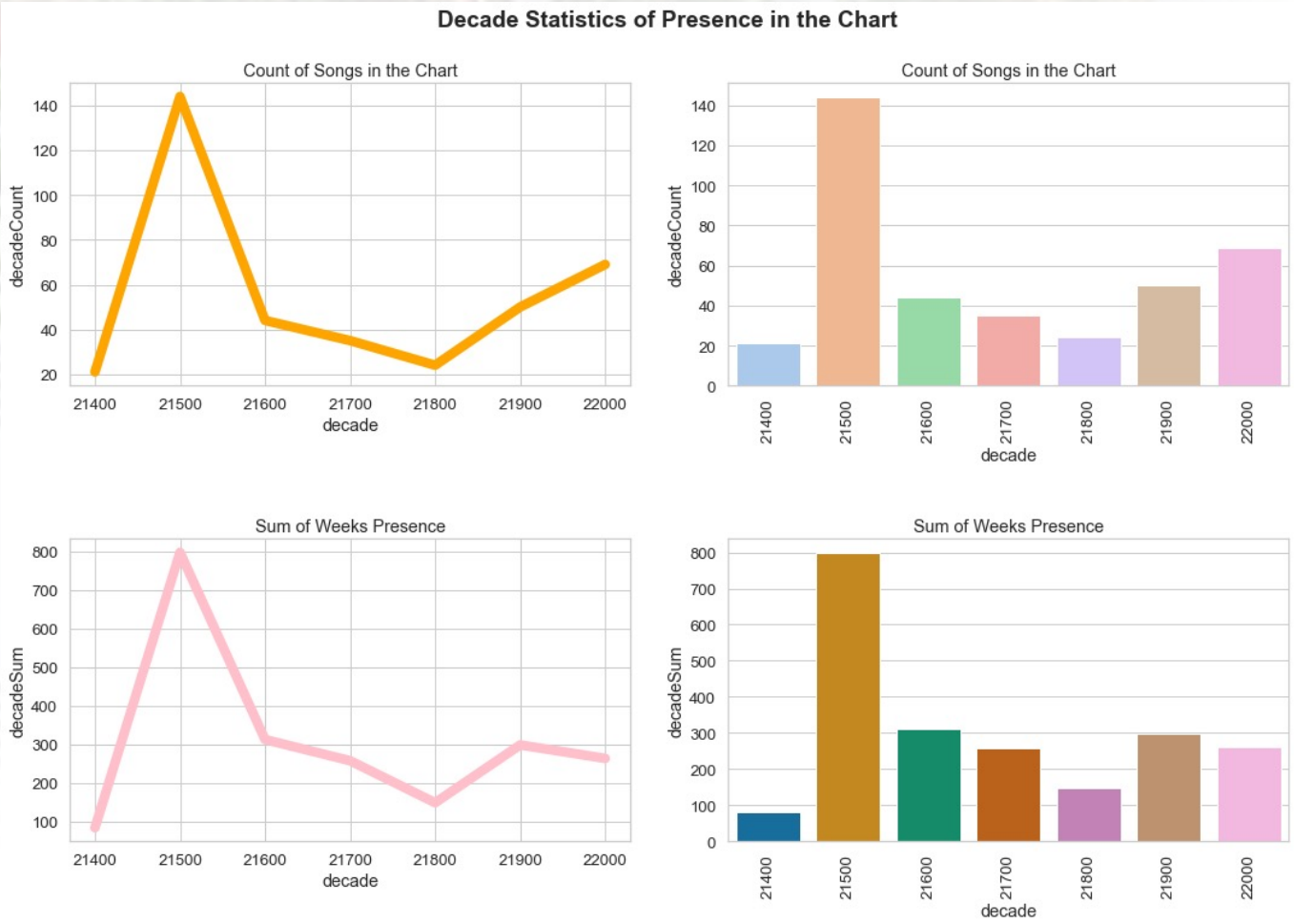
Some statistics on song presence



What can you tell me about weeks on the chart?

Some basic statistics related to decades:

It seems like Christmas-Songs were the most popular in the 1960ies while the least in 1990ies with an increasing trend in the last decade (only in terms of number of songs but not of presence in the chart!).



What can you tell me about weeks on the chart?

Machine Learning Model to predict the duration on chart

Run a machine learning model to investigate whether it is appropriate to predict weeks on chart per season and check which factor has how much influence. Because there are 13 different classes of week-duration (1-13) and too few datasets reclassify the duration as:

- presence of 1 week only
- presence of 1 month (2-4 weeks)
- presence of more than a month (5 weeks or more)

What can you tell me about weeks on the chart?

Machine Learning Model to predict the duration on chart

```
[[ 5  7  2]
 [ 0 38 10]
 [ 0  8 58]]
```

	precision	recall	f1-score	support
0	1.00	0.36	0.53	14
1	0.72	0.79	0.75	48
2	0.83	0.88	0.85	66
accuracy			0.79	128
macro avg	0.85	0.68	0.71	128
weighted avg	0.81	0.79	0.78	128

Interpretation:

- The model is 81% accurate (weighted avg)
- There is 100% accuracy for songs that are only 1 week present in the charts; for 2-4 weeks 72% and for more 83%
- 5 songs were correctly predicted as 1-week-presence-songs, 7 were erroneously classified as 2-4-week-presence-songs and 2 erroneously as more-than-a-month-songs
- 38 songs were correctly predicted as 2-4-week-presence-songs, 0 were erroneously classified as 1-week-presence-songs and 10 erroneously as more-than-a-month-songs
- 58 songs were correctly predicted as more-than-a-month-songs, 0 were erroneously classified as 1-week-presence-songs and 8 erroneously as 2-4-week-presence-songs

What can you tell me about weeks on the chart?

Machine Learning Model to predict the duration on chart

Hyperparameter Tuning – *Although many different parameter changes were tried, the model could not be improved.*

```
[[ 4  7  3]
 [ 0 40  8]
 [ 0  9 57]]
```

	precision	recall	f1-score	support
0	1.00	0.29	0.44	14
1	0.71	0.83	0.77	48
2	0.84	0.86	0.85	66
accuracy			0.79	128
macro avg	0.85	0.66	0.69	128
weighted avg	0.81	0.79	0.78	128

What can you tell me about weeks on the chart?

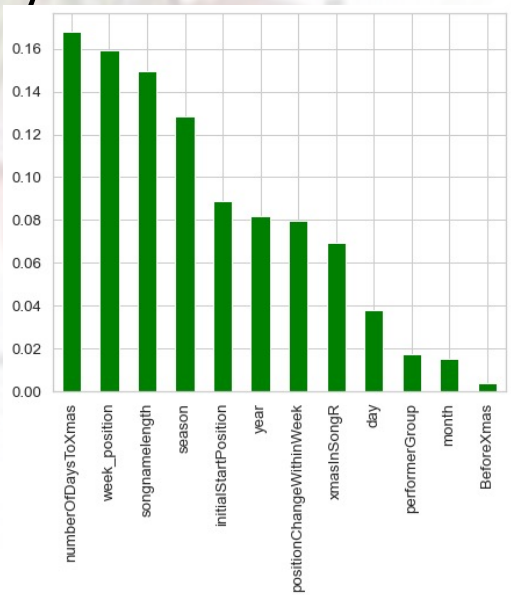
season	0.198436
week_position	0.173208
numberOfDaysToXmas	0.164639
songnamelength	0.156286
positionChangeWithinWeek	0.062544
xmasInSongR	0.061088
initialStartPosition	0.057681
year	0.055640
day	0.035147
performerGroup	0.025133
month	0.006948
BeforeXmas	0.003248

Machine Learning Model to predict the duration on chart

Feature importance

Interpretation:

- season seems to be the most influential variable with nearly 20%
- week position, year and number of Days to Christmas are also influential with each more than 10%
- the remaining 8 variables are between nearly 2% and 9%
- whether the song first appeared before or after Christmas seems to have the least impact



Some more information on the data…

Counting the numbers of datasets that have “Christmas” in their title:

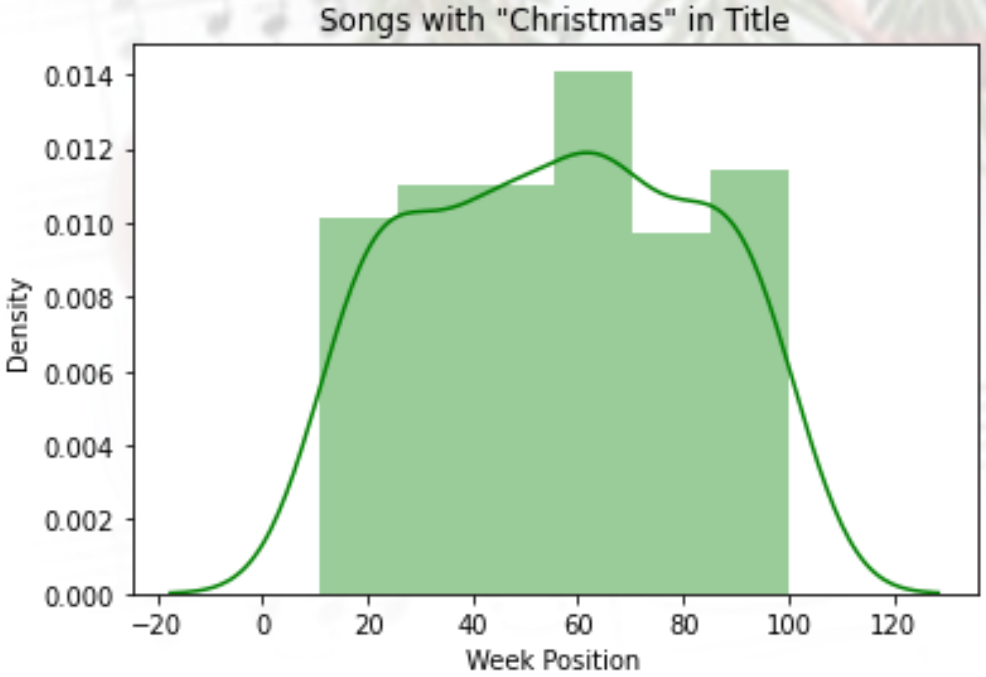
- 234 (60.5%) without “Christmas”
- 153 (39.5%) with “Christmas”

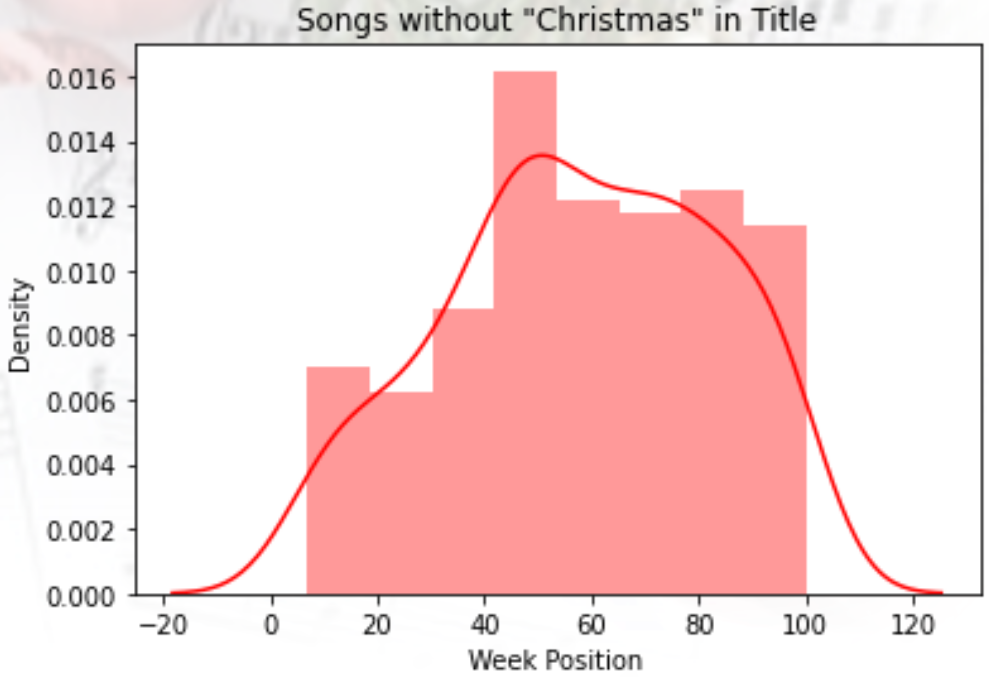
Counting the numbers of unique songs that have “Christmas” in their title:

- 44 (56.4%) without “Christmas”
- 34 (43.6%) with “Christmas”

Some more information on the data…

Some graphic information on the split:





Some more information on the data…

Some graphic information on the split:
Distribution of the records / songs: