Amanda Strack

Capstone Project I : Exploratory Data Analysis - Inferential Statistics

Are the variables significant in terms of predicting Playlist Genre?

My goal of applying inferential statistics methods to my dataset is to determine which variables are good candidates for my final model. I first explored two playlists which I found to be very similar in the data storytelling stage: Workout and Party. Next I tested the variables across all four playlists.

Comparing Workout and Party Playlists

In the data storytelling stage, we found that the variable distributions of Workout and Party Playlists were very similar. Is there a significant difference between the two genres? To test this, I used a two sample t-test. The t-test tests whether the means of two samples are significantly different.

Two sample T test at 5% significance level:

H0: The means of workout and party playlist samples are the same.

H1: The means of workout and party playlist samples are not the same.

The table below shows variables that did not result in a significant difference. All other variables resulted in a very low p-values.

Results:

	T-Statistic	P Value	Conclusion
Liveness	0.322	≈0.748	Do not reject the null, means are the same
Speechiness	1.177	≈0.239	Do not reject the null, means are the same
Duration	.903	≈0.367	Do not reject the null, means are the same

Since these variables are so similar between Workout and Party playlists, they may be weak predictors in our final model.

The t-test is only valid for continuous variables. I used the Chi-Squared test for testing our three categorical variables (mode, key, time signature). Chi-squared tests whether there is a relationship between two categorical variables.

Chi-Squared test at 5% significance level:

H0: there is no relationship between the variable and playlist genre

H1: there is a relationship between the variable and playlist genre

Results:

	Chi-Squared	P Value	Conclusion
Mode	2.408	≈0.1207	Do not reject the null, there is no relationship
Key	12.348	≈0.3381	Do not reject the null, there is no relationship
Time Signature	1.834	≈0.6076	Do not reject the null, there is no relationship

The test results tell us to accept the null hypothesis that these variables do not have a significant relationship to the playlist genre. However, it is important to check that the expected cell frequencies are greater than or equal to 5, as this is one of the assumptions for the chi-squared test. After interpreting the results, I found that time signature does not fit this assumption.

```
time signature
--------------------
Chi-Squared: 1.834
P-Value: 0.607554
Expected: [[  1.03771849   0.96228151]
 [  8.82060718   8.17939282]
 [547.91536339 508.08463661]
 [  6.22631095   5.77368905]]
```

The first two expected frequencies (1.0377 and 0.9623) are not greater than 5. Small expected frequencies is a result of the small sample sizes which you can see in the contingency table below.

Observed Values:

Time Signature	Party	Workout	Total
1	1	1	2
3	7	10	17
4	548	508	1056
5	8	4	12

To adjust for this, I will test only one category of time signature against all the others combined. Since time signature '4' is the most frequent value, I will test it against all other values to see if the category has an influence on playlist genre. A pandas method is used to create a dummy variable for only the time signature ‘4’ which equals to ‘1’ if it belongs to that category and ‘0’ if it does not.

Adjusted Observed Values:

"4" Time Signature	Party	Workout
0	16	15
1	548	508

With the new contingency table we get the following results:

```
time signature
--------------------
Chi-Squared: 0.023
P-Value: 0.879593
Expected: [[ 16.08463661  14.91536339]
 [547.91536339 508.08463661]]
The samples are independent (do not reject null hypothesis)
```

Now all of the expected frequencies are greater than 5 and the chi-squared results can be trusted. We accept the null hypothesis that there is no relationship between having a 4-beat time signature and belonging to a workout or party playlist.

In conclusion, liveness, speechiness, duration, mode, key and time signature did not test to have a significant difference or relationship between workout and party playlists. However, it is still indeterminate the amount of impact these variables will have on the final model. We will keep this variables in mind when testing different variations of the final model.

Comparing All Playlists

Next i tested the relationship of the variables to all four playlist genres. For testing the continuous variables, I used the Kruskal-Wallis test. The Kruskal-Wallis test assesses for significant differences on a continuous dependent variable by a categorical independent variable (with two or more groups).

Kruskal-Wallis test at 5% significance level:

H0: the distributions of all categories are equal.

H1: the distributions of one or more categories are not equal.

For all ten numerical variables, the Kruskal-Wallis test results in a low p-value (0.000000). Therefore, we can conclude that for each of the variables, the distributions of at least one playlist is significantly different from the others. However, the test does not identify where the differences occur.

Again, I used the Chi-Squared test for testing the three categorical variables (mode, key, time signature). The results of these tests showed a low p value for all three variables. As expected from our earlier chi-squared test, the results of time signature have expected frequency values less than 5 and we cannot trust the results of the chi-squared test in this case.

```
time signature
--------------------
Chi-Squared: 200.912
P-Value: 0.000000
Expected: [[  4.74778967   4.64169381   4.98650535   4.62401117]
 [ 37.48255002  36.64495114  39.36714751  36.50535133]
 [483.27501163 472.47557003 507.57375523 470.6756631 ]
 [ 11.49464867  11.23778502  12.0725919   11.19497441]]
The samples are dependent (reject null hypothesis)
```

As before, I will test time signature '4' against all other values to see if there is an influence on playlist genre.

Adjusted Observed Values:

“4” Time Signature	**chill**	**focus**	**party**	**workout**
0	**52**	**132**	**16**	**15**
1	**485**	**393**	**548**	**508**

With the new contingency table we get the following results:

```
time signature
--------------------
Chi-Squared: 195.453
P-Value: 0.000000
Expected: [[ 53.72498837  52.52442997  56.42624477  52.3243369 ]
 [483.27501163 472.47557003 507.57375523 470.6756631 ]]
The samples are dependent (reject null hypothesis)
```

Now all of the expected frequencies are greater than 5 and the chi-squared results can be trusted. We can reject the null hypothesis and conclude that there is relationship between having the 4-beat time signature and playlist genre.

There is a significant relationship between these categorical variables and playlist genre. However, since Chi-squared is an omnibus test, we don’t know what this relationship is but we do know that these variables are not independent of each other.