Part 1: Predictive Scenario Analysis

OBJECTIVE: 1) Use predictive scenarios to determine which influencers most affect the danceability score assigned to a song. 2) Use this information to predict the range of popularity to danceability of songs by year. 3) Compare this prediction to the actual data.

ACTIVITIES:

- Find, clean, and import data
- Create a regression analysis model
- Use these models to train two sets of predictive models
- Apply the more accurate predictive model to the dataset
- Compare the predictive values to the actual values

SOFTWARE PREREQUISITES

- SAP Analytics Cloud
- Microsoft Excel

DATA SET: Spotify - Generic Data

Scenario:
We are Bustin Jieber's management team and are partnering with Spotify on our next song release. We want to use data analytics to help guide our next song choice. Spotify's team has just sent a large amount of generic data for the past 50 years. This data lists the most popular songs by year as well as some defining characteristics like Energy, Valence, BPM, and Loudness. Each of these unique characteristics affects an overall Danceability score assigned to each song - Danceability is defined as how suitable a track is for dancing based on a combination of musical elements.

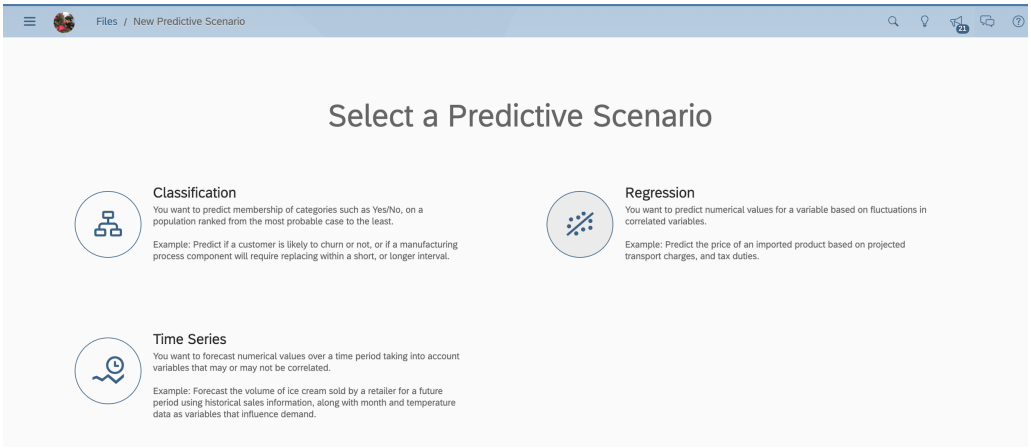
Bustin Jieber has decided he wants a hit song that will get everyone moving, so we want to determine which influencers most affect overall Danceability using Regression Analysis. Once we identify the top influencers, we want to compare the predicted values to the actual values to validate the accuracy of our data. Lastly, we want to visualize this information year over year so we can quickly understand the information and identify trends.

1. Open the Data Set
    a. For convenience, we have already cleaned the original data source and uploaded it to the SAP Analytics Cloud (SAC).

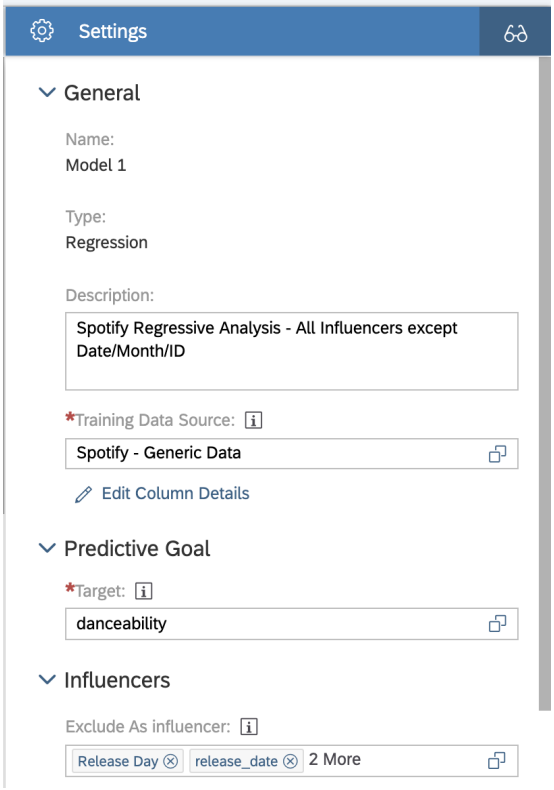
i. Split Columns out to segment year, month, day instead of dd-mm-yy
ii. Segmented data into 1990-present

b. Open the Data Sat via your web browser through SAC: https://higher-education.us10.sapanalytics.cloud/ and login
c. Open the file “Spotify - Generic Data” to your computer from the path: Cal State LA > virginiaf > group 1 data > Spotify - Generic Data

2. Create The Regression Analysis Model
   a. The goal of this model is to look at how different influencers affect danceability scores. Then to use this information to predict how danceable a certain year will be. To determine this information, we must create comparative Predictive Scenarios
   b. Use the hamburger menu in the top left and select Create -> Predictive Scenario
   c. Select Regression



   d. Choose a file path to save your Model. We stored ours in Cal State LA > virginiaf > group 1 data > Spotify Regression Model
   e. This will open your model, which is now ready for training.
   f. Under settings:
      i. Add a Description that will help you identify your Model
      ii. Select your Data Source (Spotify - Generic Data) from your path
      iii. For the first model, select all influencers under the Replicated Columns EXCEPT release day and any duplicative columns
      iv. The predictive value will be Danceability:

v. 

vi. Select Train

We will come back to these results shortly, but first let's create a comparison Predictive Model.

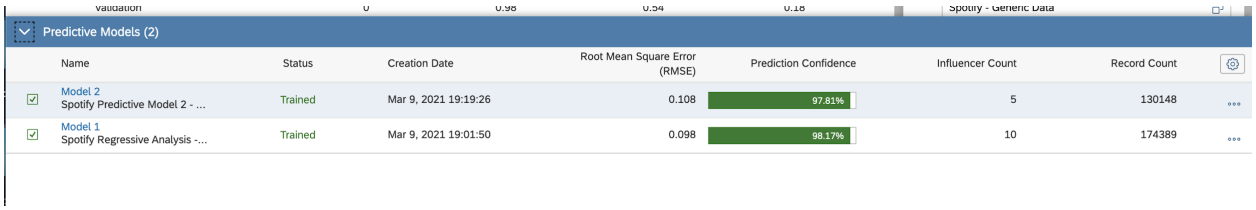
g. Create another Predictive Model using the + at the top left
h. This time, under settings:
    i. Add a Description that will help you identify your Model
    ii. Select your Data Source (Spotify - Generic Data) from your path
    iii. For this model, we will use the slider at the bottom of the Settings pop-out to limit the number of influencers to 5.
    iv. The predictive value will still be Danceability:

3. Examine the Results

You will have two models to compare now:

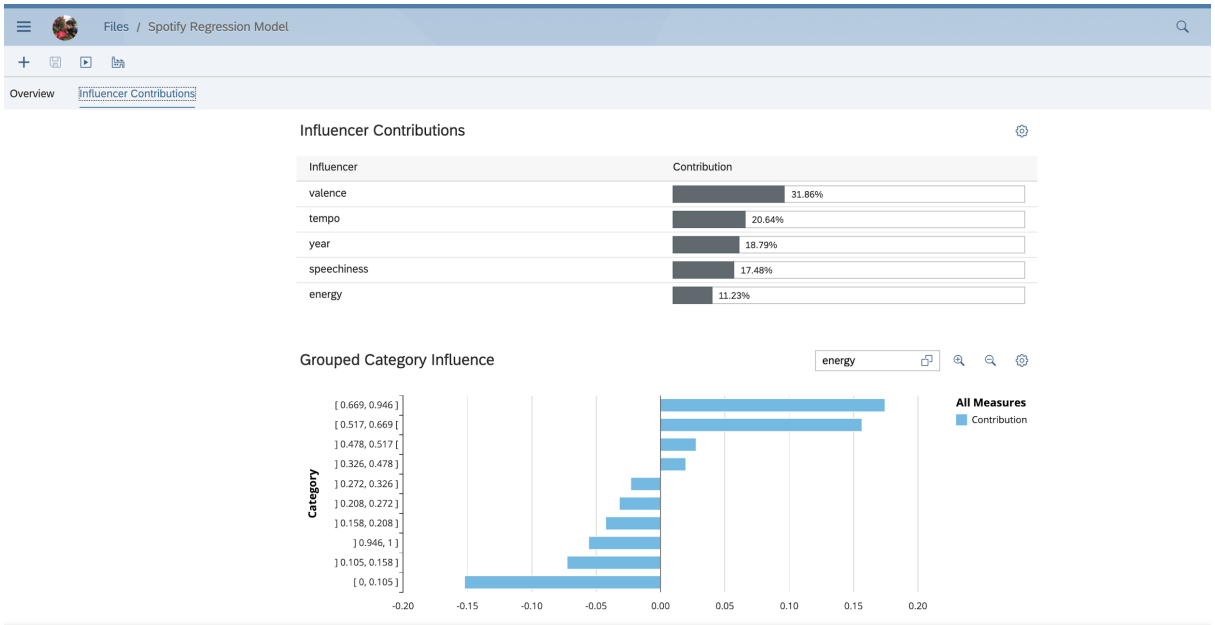
Predictive Model 1: Set influencers to all except some specific time data
Predictive Model 2: Limited influencers to just 5

Swap between the models using the blue ribbon at the bottom of the page:

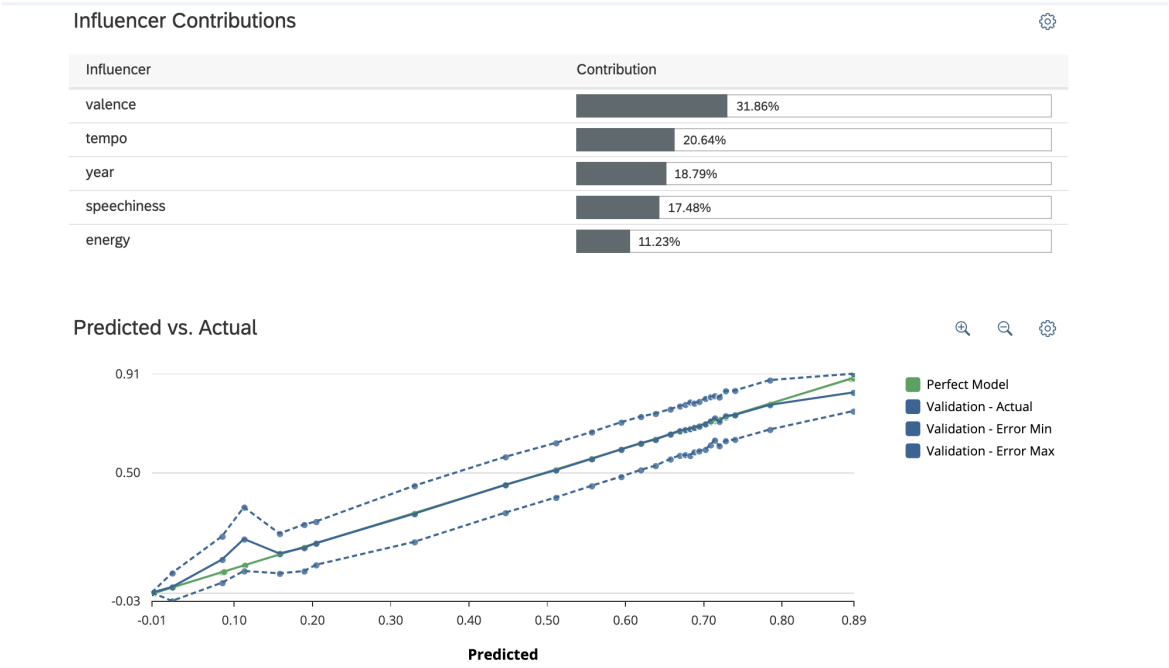


Examine the Influencer Contributions tab for both models:

Model 1:



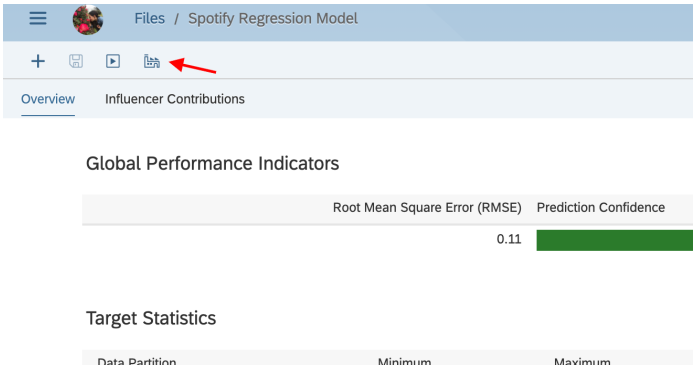
Model 2:



In Predictive Model 1 we can see that Valance is the highest influencer (24.93%). The accuracy of this prediction is 98.17%. In Predictive Model 2, we can see that Valence is a bit higher (31.86%). The accuracy of this prediction is 97.81. This shows us that the top 5 influencers carry a significant amount of weight in determining overall Danceability.

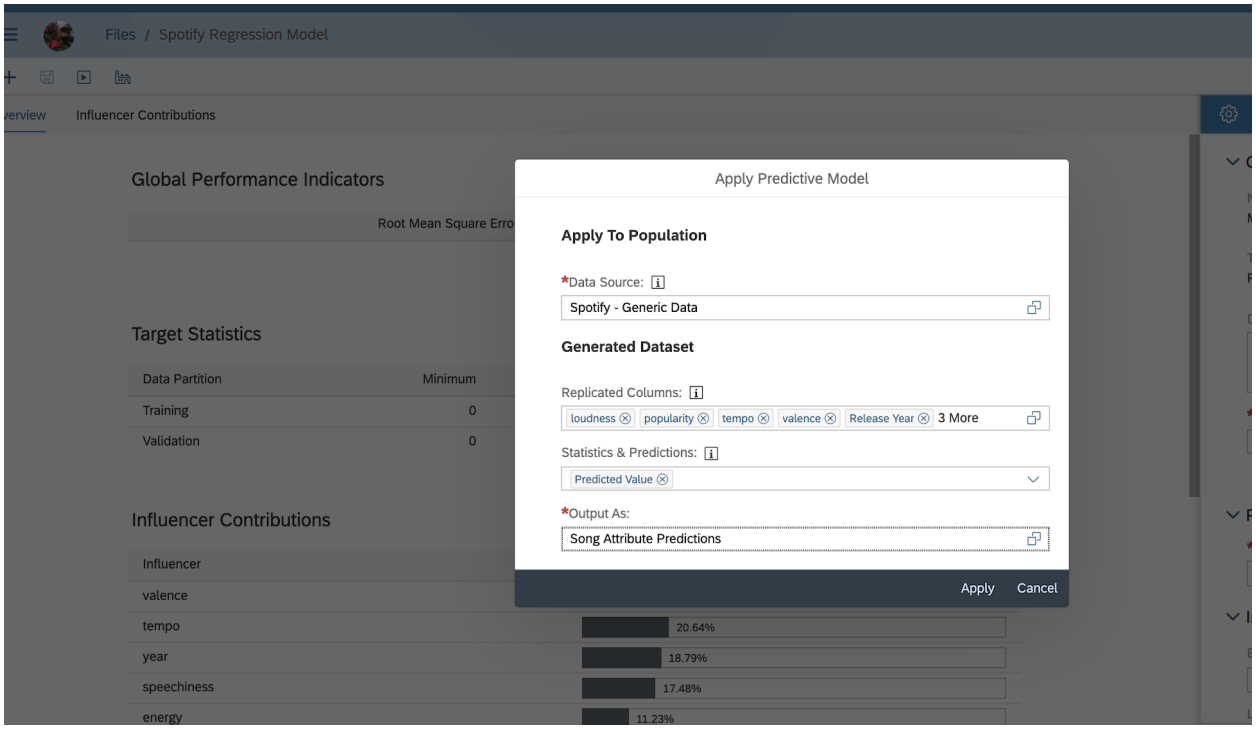
Definition: Valence describes the musical positiveness conveyed by a track. Tracks with high **valence** sound more positive (happy, cheerful, euphoric), while tracks with low **valence** sound more negative (sad, depressed, angry).

4. Apply the Predictive Model
    a. Select the model you want to use to apply the prediction
    b. Click Apply Predictive Model using the engine icon on the top left



    c.

d. This will create a new dataset with a column based on our Regression Analysis. We can use this new dataset to create a story which would estimate dancibility by year
e. Export your Data with the following Settings
    i. Data Source - Original Source (Spotify - Generic Data)
    ii. Replicate the columns. We selected the top 7 influencers:



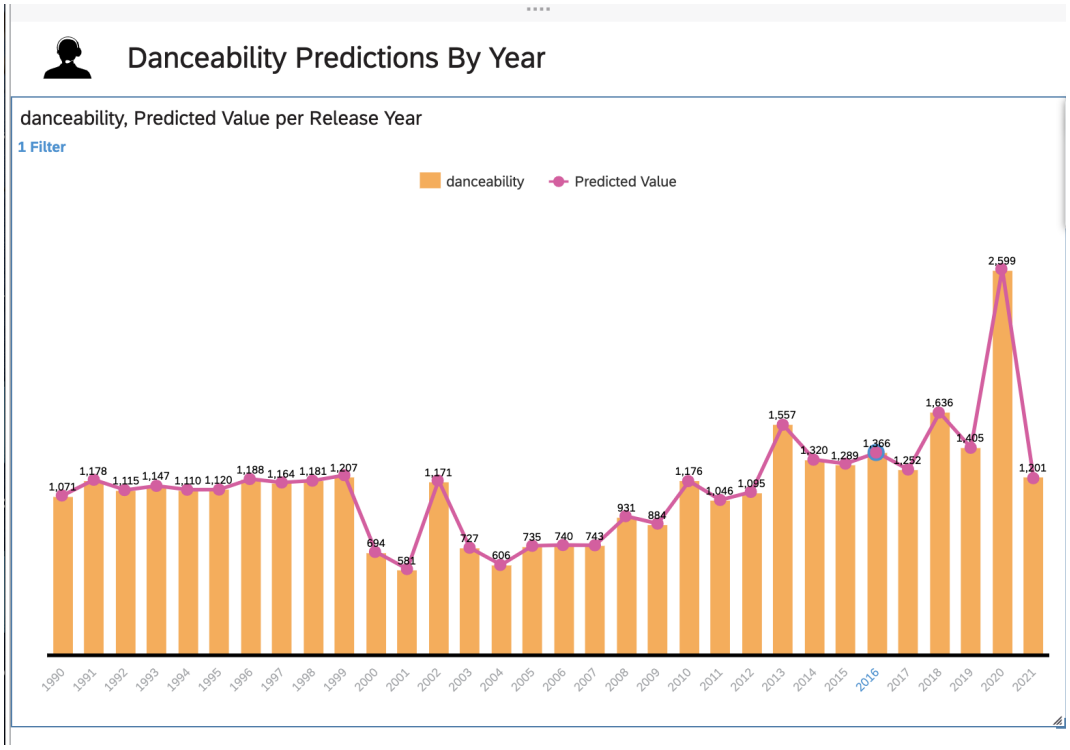
    iii. Add Predicted Value under Statistics & Predictions
    iv. Output into your folder with a descriptive name. We chose “Song Attribute Predictions”
    v. Select Apply
        1. To see the results, open a new instance of SAC and locate the output file name in your specific file path.
        2. Open you new data set
        3. Within the new data will be a column titled “Predicted Value” - this is the result of your analysis and hard work!

5. Create a story from your predictions

For our story, we want to create a visual representation that demonstrates the accuracy of our predictions and validates the importance of the influencers we have identified as the most important for determining Danceability.

a. Go to Create -> Story -> Access & Explore Data -> From an Existing Data Set
b. Select the new data set with the Predicted Values (Song Attribute Predictions)
c. Add a new chart in the Story

i. Select a Comparison Combination Column and Line chart that differentiates between the actual danceability and the Predicted Value.
ii. Segment the Dimensions by Release Year.
iii. The results are very close! This validates our predictions!



6. Consider the results

This exercise taught us a few key things:

1. The most important influencers for determining the overall danceability of a song
2. How accurate these influencers are for predicting future scores
3. How the Danceability of songs has increased over the past decade
4. Trends in Danceability over the years

Let's consider this information the next time we are enjoying our favorite songs!

Part 2: Danceability and Popularity Time-series Analysis

Objective: We created times series analysis to review and identify the trends in popularity and danceability. We created two time-series charts leveraging the dataset Average Spotify by Year.

ACTIVITIES:

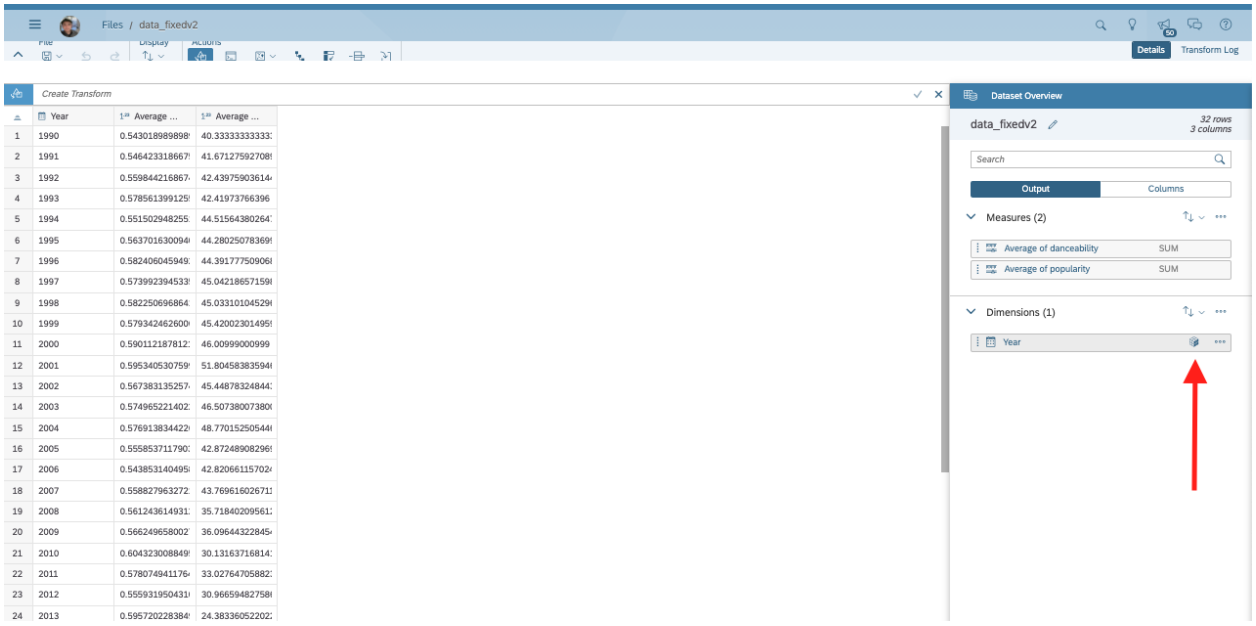
- Import new data set
- Create new page with Time series models
- Leverage automatic forecast and triple exponential smoothing

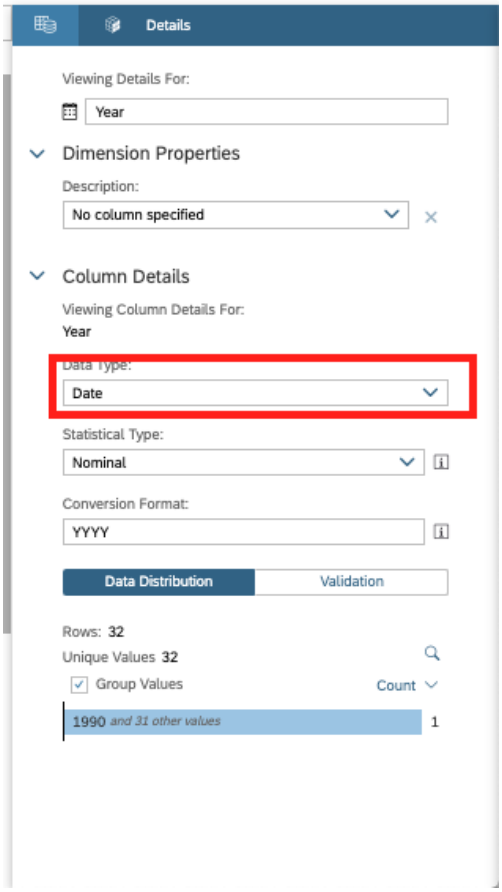
SOFTWARE PREREQUISITES

- SAP Analytics Cloud
- Microsoft Excel
- Leverage Story created “Danceability Predictions by Year”

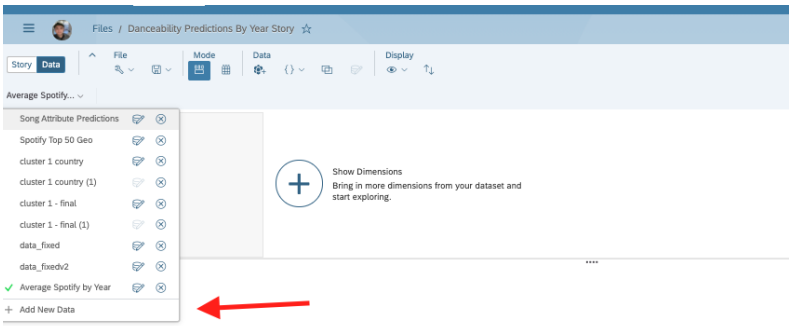
DATA SET: Average Spotify by Year

1. Before you create a page for time series, you must upload the dataset with proper data for your trend analysis.
    a. Go to create> dataset
    b. Select “data upload from file” > select source file you downloaded earlier labeled Average Spotify by Year > select **data_fixedv2** sheet. Ensure column headers box is selected.
    c. Find your existing Spotify folder and name it “Average Spotify by Year” and add **time series** as description.
    d. Once loaded, change the dimension to year by clicking on the rubix cube icon.
    e. Change viewing details to Year
        i. Date Type to “Date”
        ii. Conversion Format to “YYYY”
        iii. Rename column to Year
        iv. Remove rows 33 (Blank) and 34 (avg)
        v. Save dataset

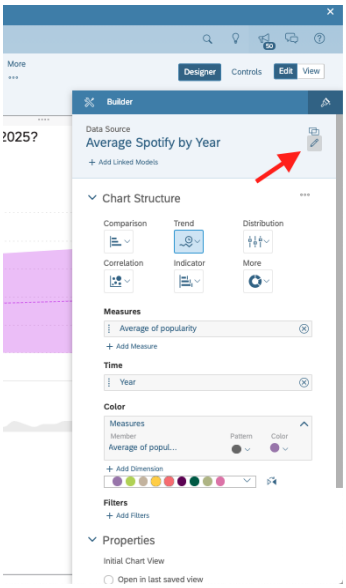

Files / data_fixedv2
Details
Transform Log
Create Transform
Year
Average ...
Average ...
1 1990 0.543018989898 40.33333333333
2 1991 0.546423318667 41.67127592708
3 1992 0.559844216867 42.43975903614
4 1993 0.578561399125 42.41973766396
5 1994 0.551502948255 44.51564380264
6 1995 0.563701630094 44.28025078369
7 1996 0.582406045949 44.39177750906
8 1997 0.573992394533 45.04218657159
9 1998 0.582250696864 45.03310104529
10 1999 0.579342462600 45.42002301495
11 2000 0.590112187812 46.00999000999
12 2001 0.595340530759 51.80458383594
13 2002 0.567383135257 45.44878324844
14 2003 0.574965221402 46.50738007380
15 2004 0.576913834422 48.77015250544
16 2005 0.555853711790 42.87248908296
17 2006 0.543853140495 42.82066115702
18 2007 0.558827963272 43.76961602671
19 2008 0.561243614931 35.71840209561
20 2009 0.566249658002 36.09644322845
21 2010 0.604323008849 30.13163716814
22 2011 0.578074941176 33.02764705882
23 2012 0.555931950431 30.96659482758
24 2013 0.595720228384 24.38336052202
Dataset Overview
data_fixedv2
32 rows
3 columns
Search
Output
Columns
Measures (2)
Average of danceability SUM
Average of popularity SUM
Dimensions (1)
Year


Details
Viewing Details For:
Year
Dimension Properties
Description:
No column specified
Column Details
Viewing Column Details For:
Year
Data Type:
Date
Statistical Type:
Nominal
Conversion Format:
YYYY
Data Distribution
Validation
Rows: 32
Unique Values 32
Group Values
Count
1990 and 31 other values
1

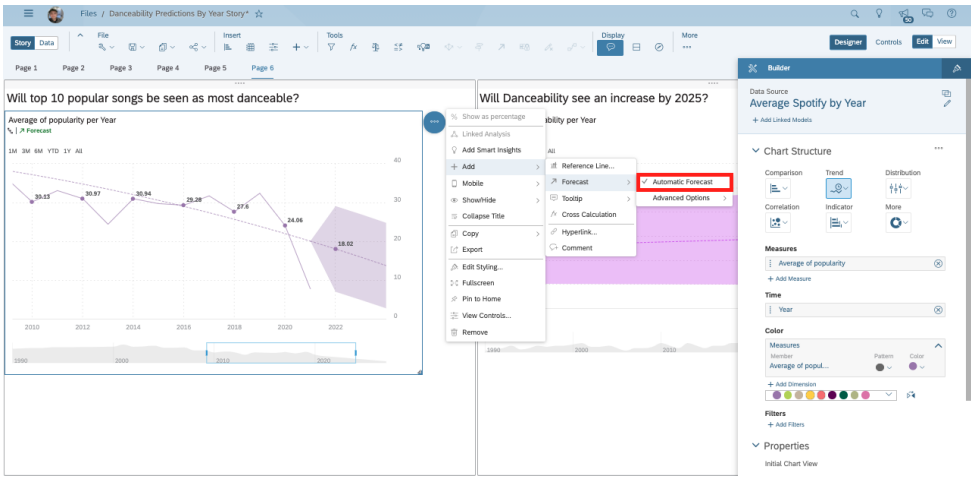
2. Now you can create another Responsive Page in the Danceability Predictions by Year story.
    a. Find your story in the aforementioned folder
    b. Create new page
    c. Add a Chart.
3. Before you add criteria, the other stories are leveraging different datasets so you need to update to reflect the right data. Add the data set “Average Spotify by Year” which can be found in the data tab, as shown in figure 3.



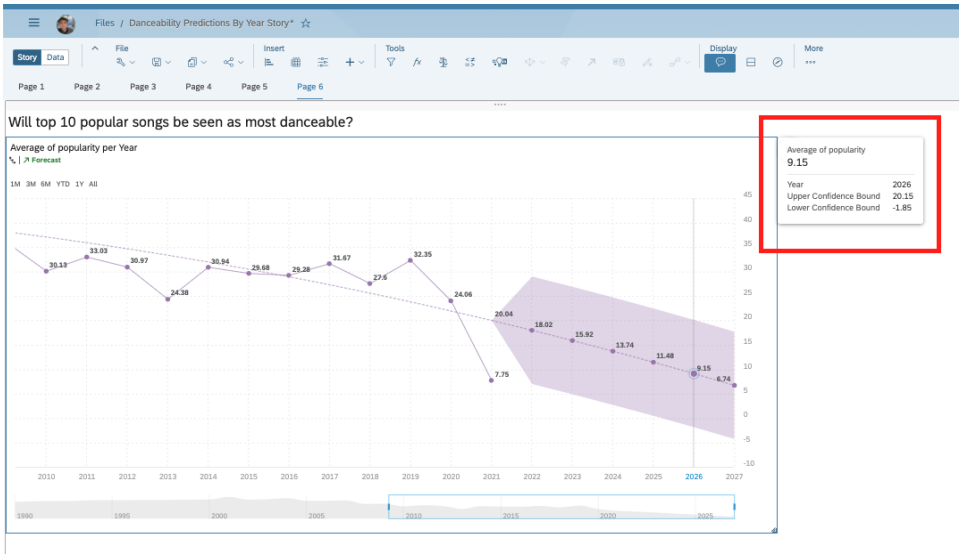
4. Switch back to the Story Tab and change the data set for your chart. Change the data source to the dataset you uploaded in the previous steps above. As shown in the figure below.



5. Use the Time Series Chart Structure listed under Trend.
    a. Add Average of Popularity for your Measure
    b. Add year as Time
    c. You can update or change the color scheme
6. This will load the chart that will showcase the trend of danceability in relation to the popularity. You can see that the danceability changes YoY but we want to forecast to help us determine which type of song we'll need to develop so you need to add a forecast:
    a. Click on the three dots and add forecast:
    b. We want to use automatic forecast
7. Add question "Will top popular songs be seen as most danceable?"
    a. Hit Save

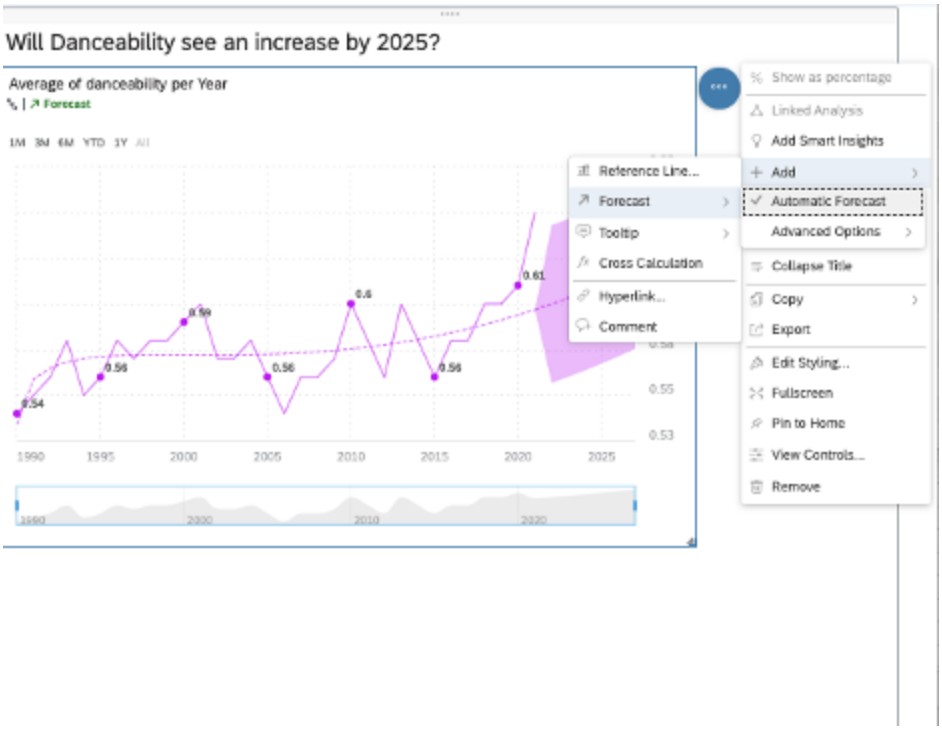


This chart is important to helping us understand how average danceability per year and popularity is changing. But we want to create a chart to help us understand if Danceability will continue to increase and whether our upcoming songs will need to be more danceable vs. popular.



The next time series chart is to analyze the shift of danceability YoY and whether that will impact our decision on the song type for our next release.

8. On the blank canvas to the right of the above chart
    a. Add chart
    b. We are leveraging the same data set so you can leave the source as above.
9. Add the details to the builder to build your chart
    a. Add **average of danceability** as your measure
    b. Add **Year** as your time dimension
    c. Feel free to change the color to a different scheme then the other chart
10. Load chart
    a. You can see that the data doesn't really tell you how danceability may potentially change over the next coming years.
    b. So we need to add a forecast
        i. Click on the three ellipsis and select automatic
11. We can tell that the chart doesn't really help us or account for seasonality so we must change the forecast.



12. Change the Forecast to Triple Exponential
13. Click Save.

You can see that by 2025 the danceability is forecasted to continue to increase to more than 10% from 2020 danceability average. Also much higher from the low of the 1990 danceability song. This indicates and can further help us determine that we may need to focus on a song with more danceability in the upcoming years.

Part 3: Geo Mapping

OBJECTIVE: The objective of this exercise is to show visualization of various popular song types in different countries, which will help Bustin Jieber's manager determine where to release certain songs and choose tour locations based on their "styles" or "type".

ACTIVITIES

- Import and prepare data.
- Create a story using responsive pages.
- Create a geo map with two layers
- Experiment with formatting, colors, and other styling components.

SOFTWARE PREREQUISITES

- SAP Analytics Cloud
- Microsoft Excel

DATA SET: Data file titled top50contry.csv

SCENARIO

The band Bustin Jieber publicist is finalizing plans for the promotion and tour for Bustin Bieber's latest album. The album is full of high energy, high BPM songs. The publicist has acquired data, in a CSV file, from Spotify that shows the popularity of songs based on the song's energy, bpm, valance and popularity scores.

The publicist wants to segment the 19 countries into three different groups by the popularity of songs based on dancibility and energy (nrgy) so that specific strategies can be applied to each country's segment. We will use Geo Map and layers to display the different potentials of each country.

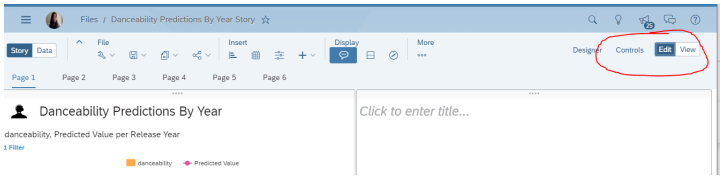
1. **Locating the dataset**
    a. Through your web browser, go to SAP Analytics Cloud (SAC):
        i. https://higher-education.us10.sapanalytics.cloud/ and login
        ii. Download the file "top50contry.csv" to your computer from the path: Cal State LA > virginiaf > group 1 data > Spotify - Generic Data

2. **Acquire Data and adding to the Story**

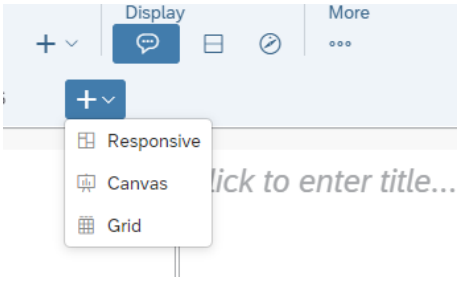
There is more than one way to start your story. This is just one of them.
Visualizations within the stories are created as either responsive pages or canvas pages. Responsive pages can be presented in various formats – they will adapt to the screen size. If you

expect to present the story on various devices, perhaps on a phone or in a boardroom with big screens, then you would want to choose Responsive Pages. If you want pixel-perfect visualizations or if you do not need the pages to adjust to the display medium, then use Canvas Pages.

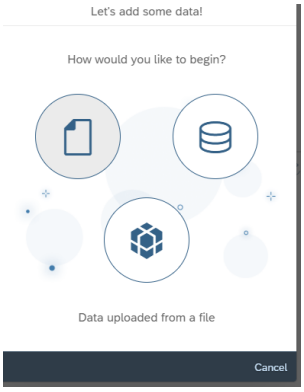
1. Add to existing story
    a. Open existing story from the path: Cal State LA > virginiaf > group 1 data > *Danceability story (still need to change)*
    b. Select “edit” button the right corner as seen below:



    c. Locate the “+” sign next to “page 1” of the story, click “+” to add a new responsive page.

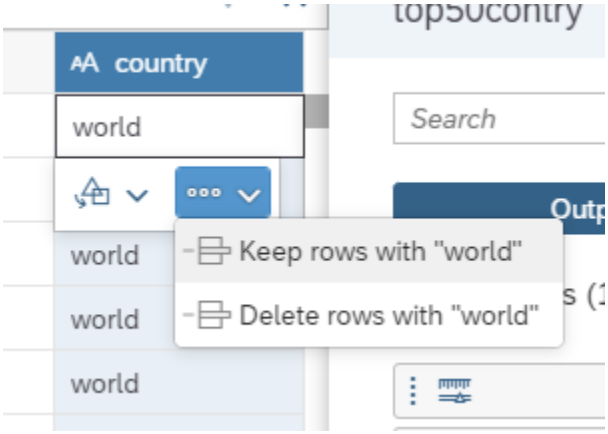


2. Explore data and align it to fit scenario
    a. Click on “Data” view
        i. from the drop down menu below this “data” button, indicated with a “v” symbol
        ii. Click “add new data” from this drop down menu
        A pop-up appears, and select the one that says “ data uploaded from a file”



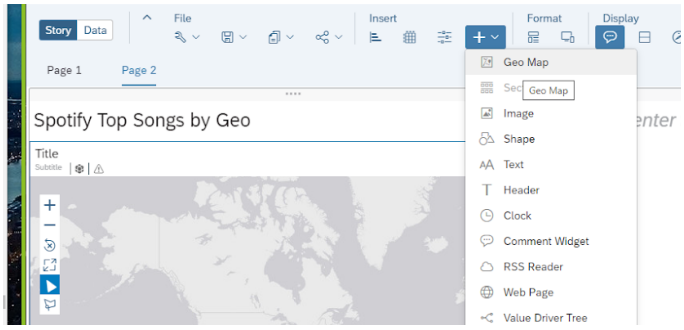
        Select source file ⟶ select the file you’ve downloaded from earlier “top50contry.csv” , use default settings with first row as column headers and CSV delimiter as auto detect.

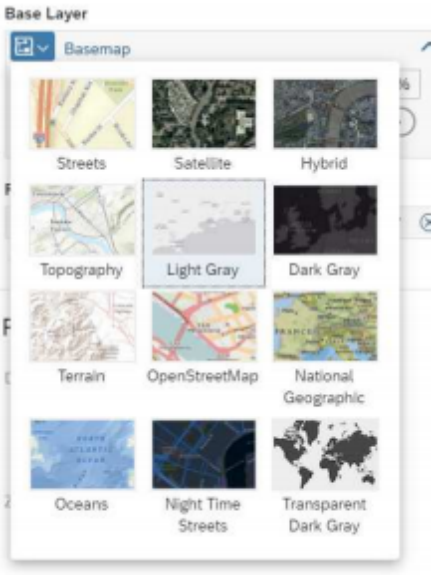
b. Click on “grid view” on the data view of the story.
    i. Locate the “country column” and delete rows and columns with “**world**”. As we only want to focus on 19 countries. Refer to the screenshot below to delete.



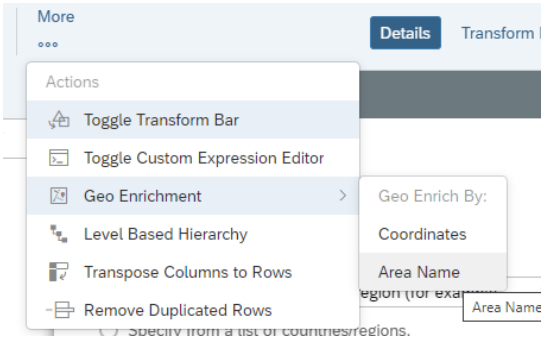
3. Create a Geo-Map with layers

a) Go to the Story Tab. (The Story Tab is in the upper left corner of Figure below)
    i. The default view for a responsive page is two lanes arranged vertically as two frames. Lanes are used to organize content on multiple devices. If you were to create a canvas page, you would not see lanes.
    ii. Each Lane can have its own style.
b) You can add or delete lanes as you see fit.
c) On one of the lanes, click to add a title “Spotify Top Songs by Geo “; See the below.
d) Insert a Geo Map by choosing the Insert + and choosing Geo Map from the drop down list of options.
e) On the Designer Builder panel to the right of the chart and lanes, choose Basemap. a. Choose the style of the background, “base map” → “Streets” for your visualization.



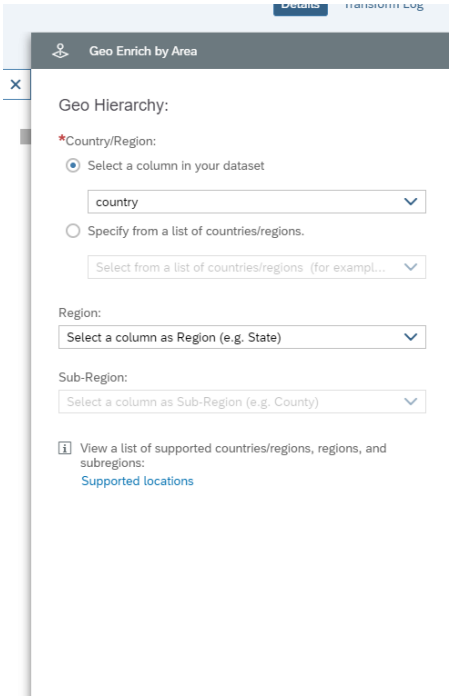


f) Go back to “data” view and click three dots under “more” menu and select “Geo Enrichment” → “ Area Name” as shown below:



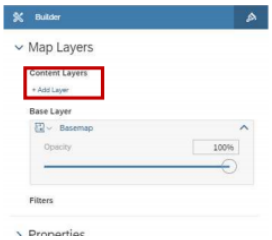
g) A Geo Hierarchy will pop up, and the selection below must be made:

i. Country/Region: country → click OK

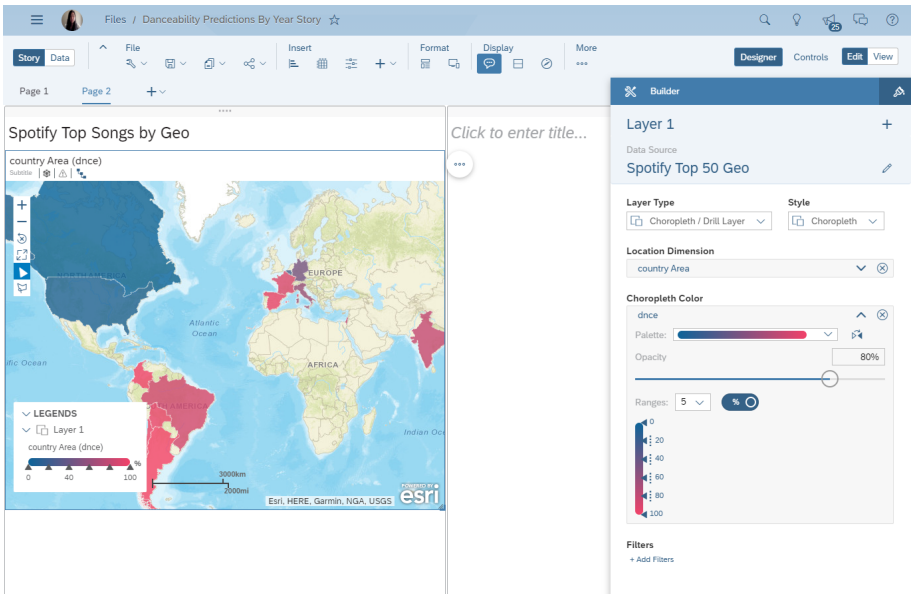


h) Go back to Story view and select on the Geo Map and click “Edit” Designer

i) + Add Layer Choropleth/Drill Layer



j) For the 1st layer , we are interested in “country” and “danceability”



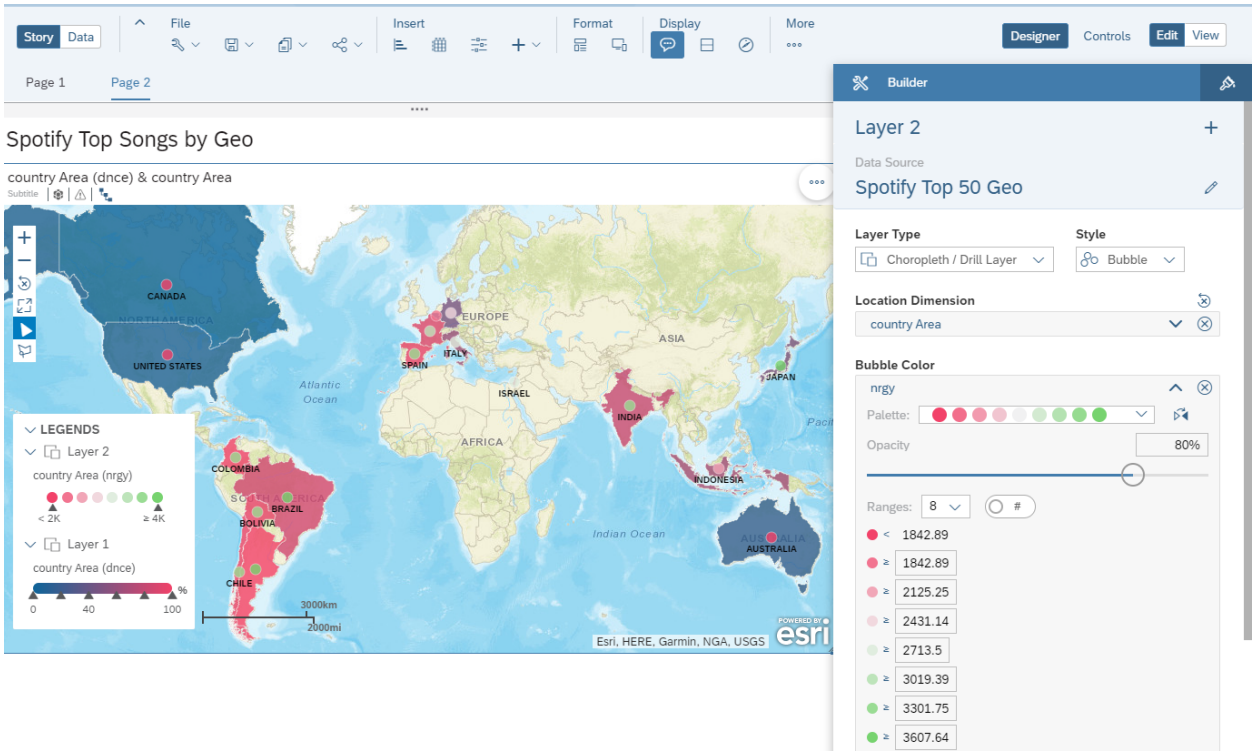
i. Under Location Dimension, choose country Area. (It should be the only choice.

ii. Under Choropleth Color, choose dnce for “danceability”.

iii. choose Range: 5

k) + Add Layer Choropleth/Drill Layer ⟶ style: “Bubble”

l) For the 2nd layer , we are interested in “country” and “energy”



i. Under Location Dimension, choose country Area. (It should be the only choice.

ii. Under Bubble Color, choose nrgy for “energy”. Choose a different color palette of your choice to help you distinguish.

iii. choose Range: 8

m) Be sure to click save to save your updated story.

Part 4: Segmenting Countries Using Clustering SAC

OBJECTIVE: The objective of this exercise is to segment countries based on various attributes to help plan with Bustin Jieber's new song promotions.

ACTIVITIES

- Import and prepare data.
- Apply Smart Grouping cluster analysis.
- Create data visualizations.
- Analyze and interpret output from models.

SOFTWARE PREREQUISITES

- SAP Analytics Cloud
- Microsoft Excel

DATA SET: Data file titled top50contry.csv

SCENARIO

The band Bustin Jieber's publicist is finalizing plans for the promotion and tour for Jieber's latest album. The album is full of high energy, high BPM songs. The publicist has acquired data, in a CSV file, from Spotify that shows the popularity of songs based on the song's energy, bpm, valance and popularity scores.

The publicist wants to segment the 19 countries into three different groups by the popularity of songs based on bpm and energy (nrgy) so that specific strategies can be applied to each country's segment. We will use clustering of countries data to assist the manager in developing promotion strategies.

1. **Acquire the data**
   a) Through your web browser, go to SAP Analytics Cloud (SAC): https://higher-education.us10.sapanalytics.cloud/ and login
   b) Download the file "top50contry.csv" to your computer from the path: Cal State LA > virginiaf > group 1 data > top50contry.csv

2. **Visualize the Country data**
   a) From the main men, select Create → Story.
   b) Select Access & Explore Data.
   c) Select Data uploaded from file.

d) Search for the top50contry.csv file in your folder where you downloaded the file "top50contry.csv", Open.
    1. Use first row as column headers should be selected and the CSV Delimiter should be set to Auto-detect.
    2. Import.
    3. You will be directed to the Data view. You should have 1001 rows of data: remove all the measures, except for the following six measures (year, bpm, nrgy, popularity, dancebility, valence) and one Dimension (country) in the dataset.
e) On the story view, insert chart
f) Select Bubble Chart on the Correlation chart options. NOTE: While the option will be grayed out, you can still select it.
g) Construct the chart as follows:
    1. + Add **bpm** to the X-Axis.
    2. + Add **nrgy** (energy) to the Y-Axis.
    3. + Add **pop** (popularity) to Size.
    4. + Add **country** to the Dimensions.
    5. Add a Tooltip Measure as shown in Figure 1. Hint: From the chart structure + of the chart Builder, select + Tooltip, and then select Measure.

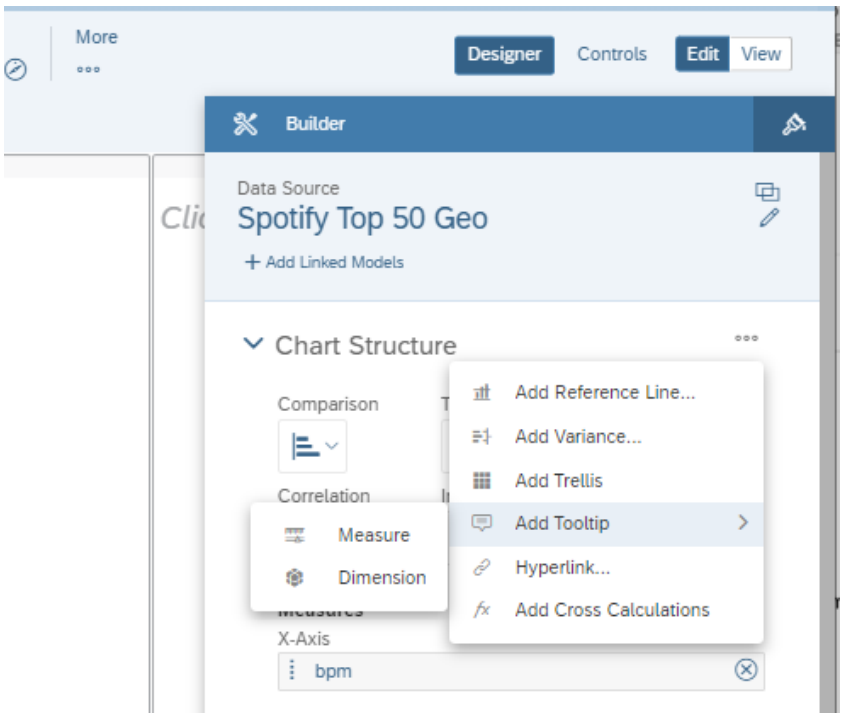


Figure 1: Adding a Tooltip

    6. Tooltip Measures will now show as a Chart Structure option. + Add Country Size to Tooltip Measures.

7. You will now see a bubble chart of the first six measures by Country.
8. Save it as “Danceability Predictions by Year Story”

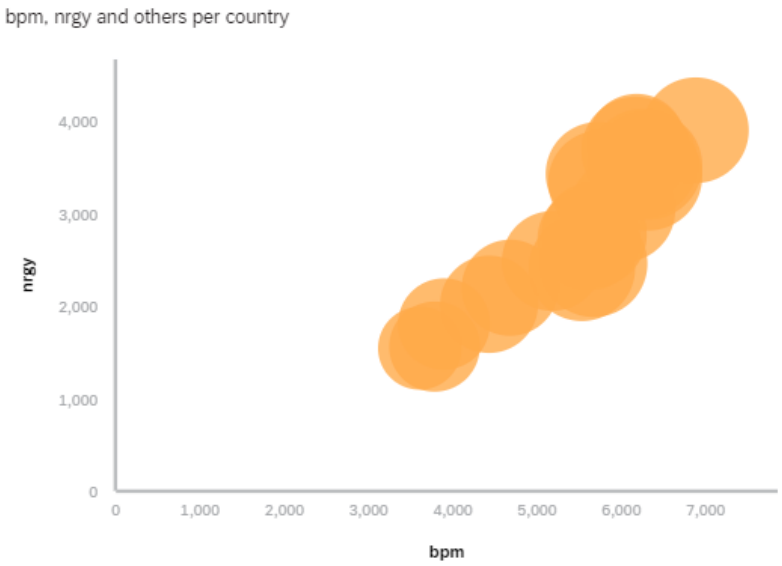


Figure 2: A Bubble Chart of Country Data

3. Creating the Cluster Analysis

Group the countries that are similar using a k-means. In SAC, clustering is done using Smart Grouping.

a) Toggle on Smart Grouping (near the bottom of the Builder panel). Hint: See the Figure 3.
b) Change the Number of Groups to 3, (3 is k in the k-means algorithm).
c) Change the Group Label to “Cluster” just to be consistent with our understanding of cluster analysis.
d) Select Include Tooltip Measures in grouping so that all seven measures are considered in the cluster analysis.

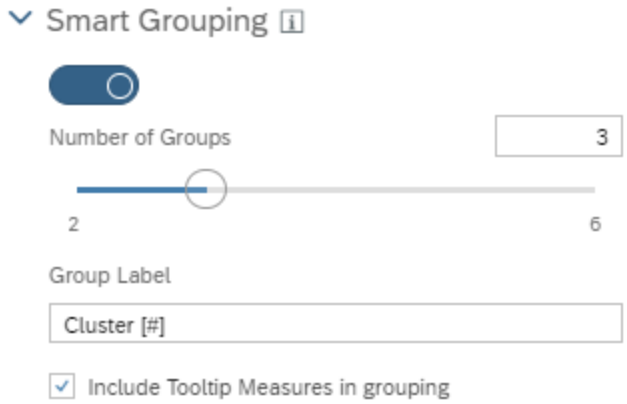


Figure 3: Configure Smart Grouping

e) You should now see three distinct groups (clusters) in your chart.
f) The clusters in the default monochromatic color scheme tend to blend together, so change the Color pallet.
g) Filter on the clusters by clicking on the cluster number you wish to examine.
h) Add “Where should Bustin Jieber release his new song?” as the title of the clustered bubble chart
i) Save the story

4. Visualization and Interpretation

The results of the grouping (clustering) can be further analyzed by associating the cluster numbers with the data in the country data “model”. (The original top50contry.csv file is stored as a private or embedded “model” within your SAC story.)

a) Merging your data sets.
    1. Filter to Cluster 1 by clicking “Cluster 1” on the legend of the bubble chart you created. Refer to Figure 4.

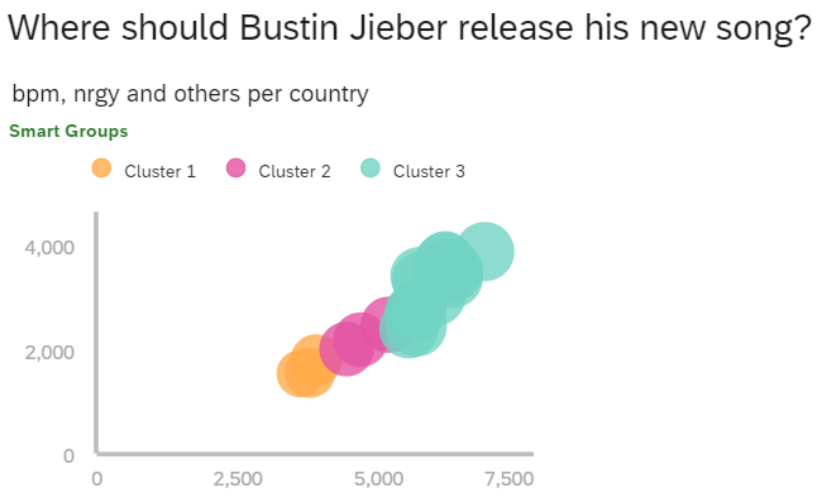


Figure 4: Filter on the Cluster

b) Click on Filter icon.

1. The SAC Smart Grouping will continue to segment the filtered data set to even smaller clusters as seen in Figure 5.
2. Export from the chart dropdown list.

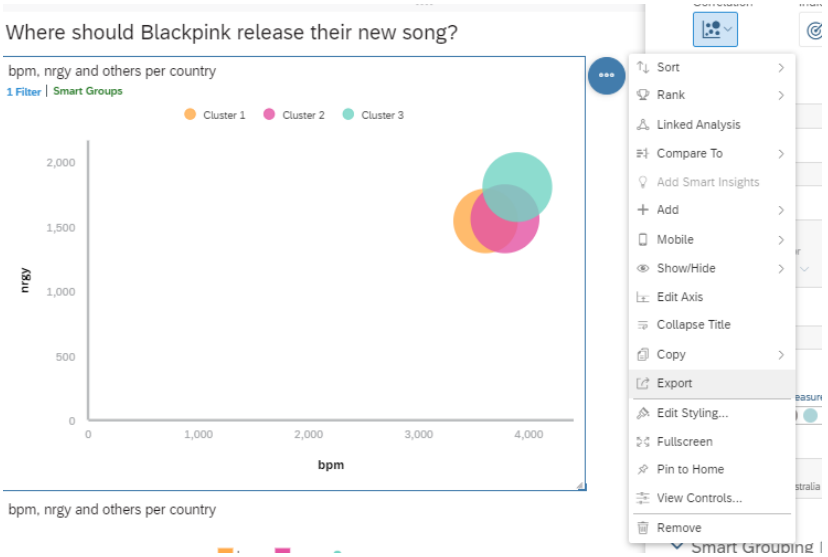
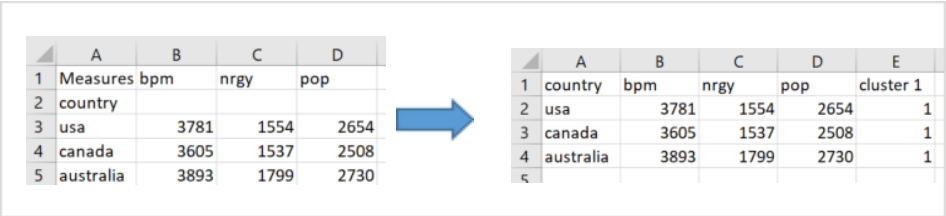


Figure 5: Export the Clustered Data

3. Save the .csv file as “Cluster 1” and on your computer. Then repeat steps for “Cluster 2” and “Cluster 3” totally three downloaded files.
4. Prepare the cluster data for integration with the Country data model in SAC.

- The first step is to clean up the header so that it is only one row. Open the Cluster 1.csv file.
- Move content of cells B1:D1 to cells B2:D2. See Figures 6 and 7.



	A	B	C	D
1	Measures	bpm	nrgy	pop
2	country			
3	usa	3781	1554	2654
4	canada	3605	1537	2508
5	australia	3893	1799	2730

	A	B	C	D	E
1	country	bpm	nrgy	pop	cluster 1
2	usa	3781	1554	2654	1
3	canada	3605	1537	2508	1
4	australia	3893	1799	2730	1

Figure 6: The Cluster1.csv file from SAC

- Then add a column called “Cluster” add “1” in each cell in the column for file “Cluster 1” and repeat these steps for clusters 2 and 3.

5. Now you need to merge this cluster data with the country data
   - Make sure you’re on the Data view, then in the form the dropdown menu next to Country, select + Add new Data.

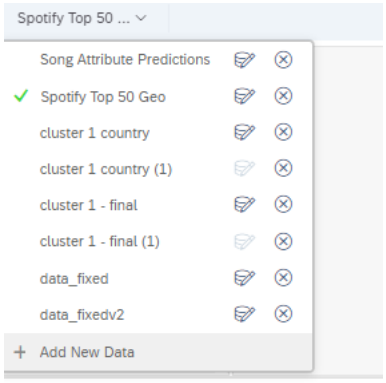


Figure 8: Adding Data to the Analysis

6. Select Data uploaded from a file.
7. Select Source Cluster 1.csv.
   Note: Use first row as header should be selected then Import.
8. As Fig 9, on the Save dropdown, select Open with Basic Data Preparation. This will allow you to append the files for clusters 2 and 3.

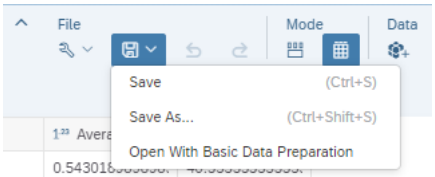
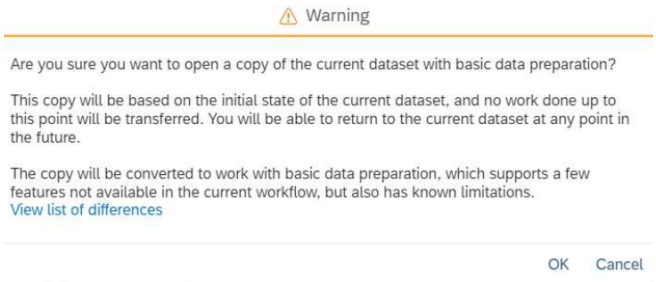


Figure 9: Open with Basic Data Preparation

(6) If you see the following warning click OK.



9. Select Reimport Data from the Data ribbon.
10. Select Cluster 2.csv by selecting “Import File” in the pop-up window.
11. With the default option, select Import.
12. When you see the following screen, select Append.

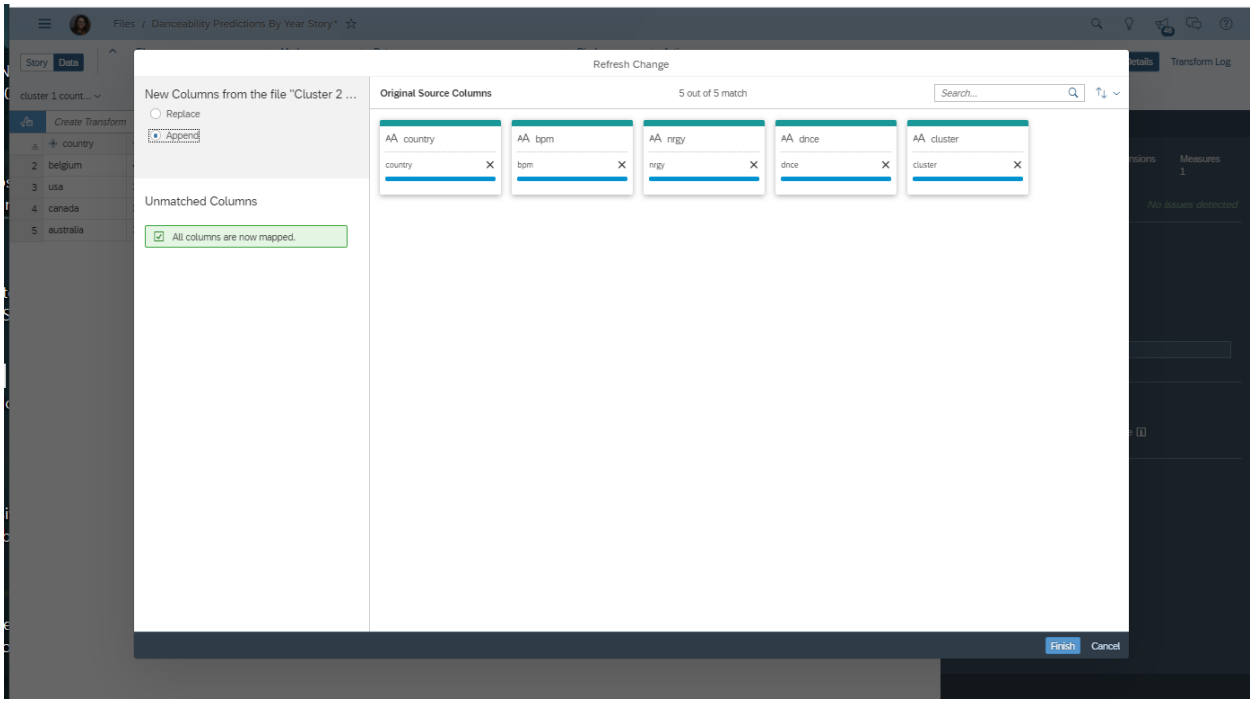


Figure 10: Append a File

13. Finish.

14. Repeat the append for Cluster 3.csv.
15. Now take a look at the data in the Clusters data set and you should find countries in all three clusters and Save.

2) You can visualize the Country and Clusters data in the Story view.
   a) Add a new page (through either a Canvas or a Responsive page. Both have a Chart menu) and add a chart.
   b) Add a calculated dimension for Countries, as shown below. The path is: Add Measures > + Create Calculated Dimension. Then select OK.

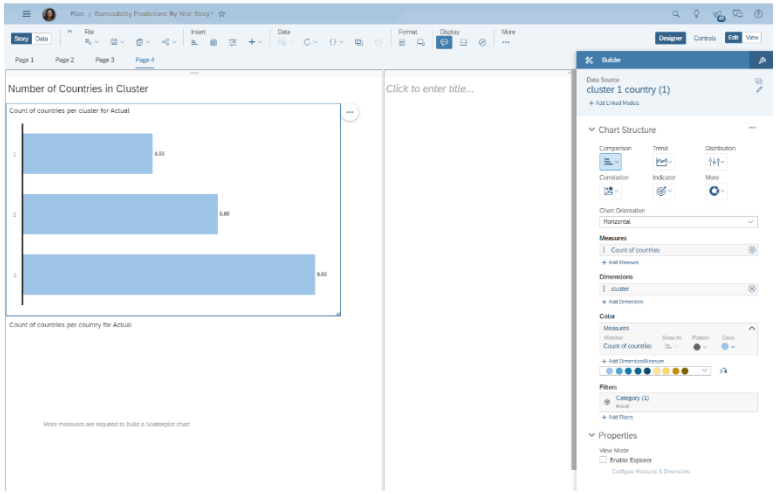
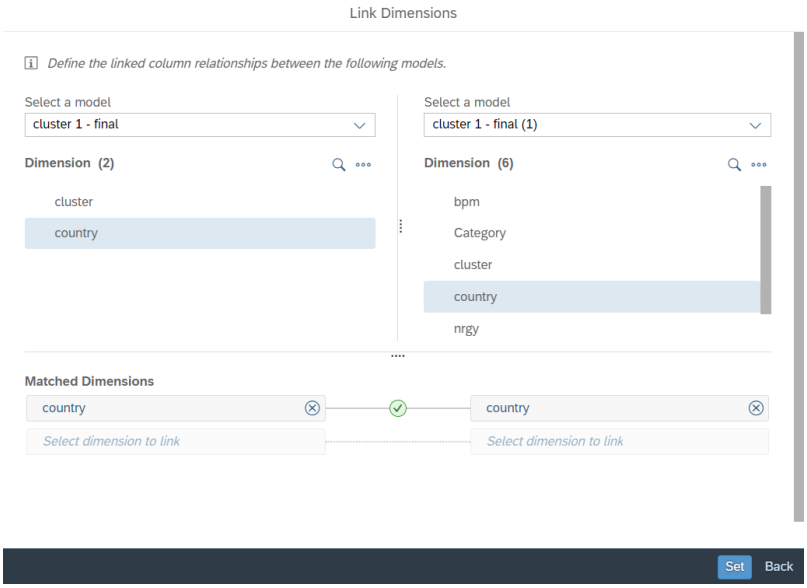


Figure 11: Count of Countries

   c) In builder, In Builder, select the Link Dimensions icon to the right of Data Source. You should see Clusters “Cluster 1 (1)” as a choice.
   d) Select Clusters “Cluster 1 (1)” by clicking on the box.
   e) You now need to choose the matching dimensions from each data set.
   f) In our case, the “countries” dimension in the countries data matches the “countries” dimension in the Cluster data.
   g) Select Set
   h) For the data merge to work, we need to also select whether we want to link based on the ID or the description of the data in the column. Our data has no description, so we need to select Data Samples > ID. The Link Dimension settings are shown below.



[insert screenshot] Figure 12: Link Dimensions

i) Click Done on the following popup.

3) Leave the chart as a column chart. To add variables to the chart, you will now have a choice of which data set you would like to use. You will see them as a drop down when you add a measure or dimension. SAC calls this a blended data chart.
   a) Add Count of Countries from the Country data set to Measures.
   b) Add Cluster from the Clusters data set to Dimensions.
   c) You will see the chart as follows. Save the story.

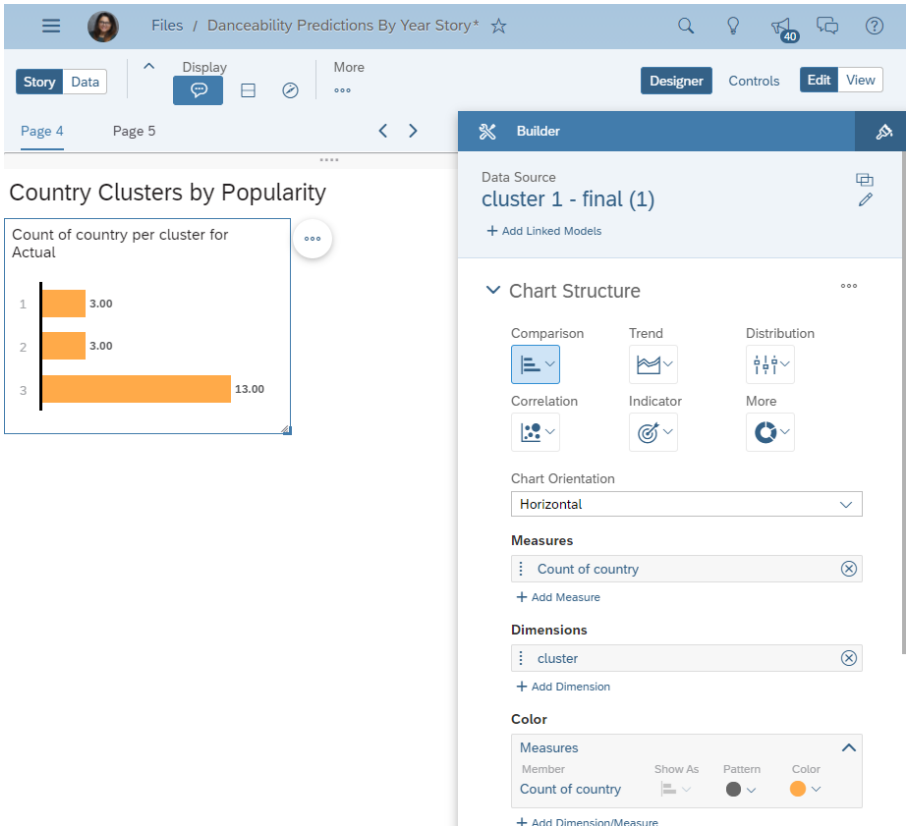


Figure 13: Countries in each cluster

4) At page 3, you can see the detail of each country in a cluster. And, at page 4, you can see how many countries are in each cluster.