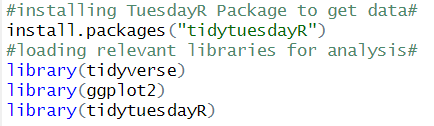
*Scooby Dooby-Doo: WHERE ARE YOU?!?!?*

The purpose of this project is to demonstrate my proficiency using R to perform exploratory data analysis, clean and wrangle data, utilize relevant statistical tests, and produce visualizations that help tell the story of the data. Moreover, I will utilize R to perform data analysis and visualization techniques. Additionally, after my analysis in R, I will transition to utilizing Tableau. In Tableau, I will further explore the world of Scooby-Doo through visual analysis. I will demonstrate my proficiency using Tableau to produce visually-appealing and insightful visualizations. To access my full dashboard, please use the following link: <https://public.tableau.com/views/VisualExplorationofScooby-DooUniverse/CulpritMotive?:language=en-US&:display_count=n&:origin=viz_share_link>.

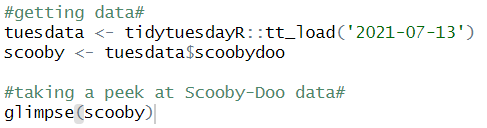
If you are not already familiar with the Scooby-Doo franchise, Scooby-doo is essentially an American media franchise based on an animated television series launched in 1969. Since then, the franchise has continued to produce several derivative media, and the franchise is still serialized to this day via a variety of mediums.

The dataset used in this case study is publicly available from Kaggle using the following link: <https://www.kaggle.com/datasets/williamschooleman/scoobydoo-complete>. I would like to thank Kaggle user Plummye for creating this dataset!

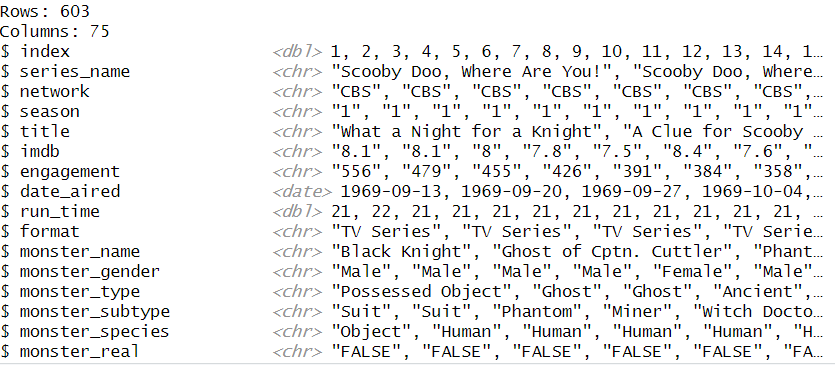
Ultimately, the first thing I want to do is load the dataset into R as well as load relevant packages that would be beneficial in my analysis.



Next, I want to actually import the dataset into R and take a peek at what is contained within the dataset.



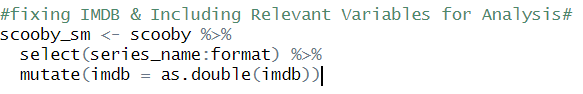
Using the glimpse command, I can see essentially what the dataset entails.



As an illustration, I can see that the dataset contains 603 rows and 75 columns. I can also see all of the variables within these columns as well, such as index, series name, network, season, title, does batman appear, which character revealed the monster, did anyone say “zoinks”, “jeepers” “jinkies” or “groovy” throughout the episode, etc. Likewise, I can see how these variables are defined. For instance, index is defined as a double variable, title is a character, and so on.

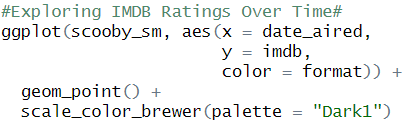
For the purposes of this analysis, what I really want to explore is IMDB scores and how the variation of IMDB scores change with respect to different characteristics. For instance, is IMDB score correlated with the network that published the cartoon? Or, is date aired correlated with IMDB scores? Etc. Answering these questions might give me some insight with respect to series quality over time by network and general interest in the series from the population over time as well. At this point in time, my analysis is relatively unstructured – I just want to explore things I think would be interesting; however, I do have some general questions in mind that I would like to answer if possible.

Regardless, I immediately notice some problems with the data. For instance, there are 75 different variables. While there are a lot of interesting stuff here, I ultimately really just want to focus on a few, tailoring my analysis accordingly. As aforementioned, I really just want to focus on IMDB ratings as well as variables that might be correlated with that. So, I next endeavor to essentially remove some of the variables that are not relevant to my analysis from the dataset. I also notice that the IMDB variable is defined as a character, whereas it really should be a double or number. So, I run the following code:

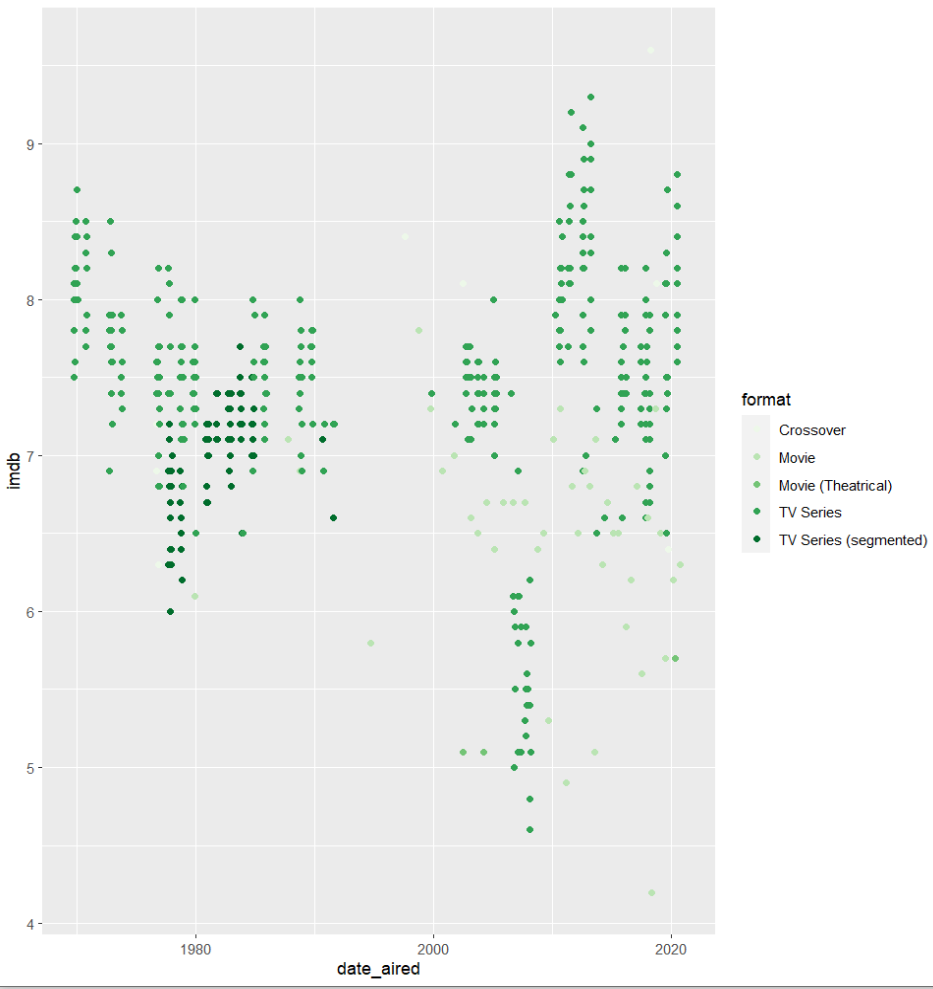


In running that code, I essentially made a new data frame containing only variables relevant to my analysis. These variables would be series name, network, season, title, IMDB, engagement, date aired, run time, and format. I also changed the variable type of IMDB from a character into a double.

Next, I just want to explore IMDB ratings over time in general. I just want to get a feel for how the data is distributed. For example, are there any outliers? Is the distribution normal or skewed? What is the general mean and median of the distribution over time? So, I want date aired to be my x-axis on the visualization I produce, and I want IMDB to be my y-axis. To explore this, I run the following code:



In running this code, I also produced the following visualization:

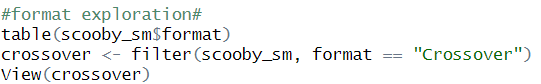


From the visual, I can get a feel for how IMDB ratings have changed over time. I can also see what format the observations are, such as movie or tv series. As an illustration, IMDB scores were relatively high when the cartoon first started airing; however, ratings declined over time up until about the year 2000. But, after the year 2000, ratings then began increasing once again somewhat all the way up to modern day. There is also more spread towards with respect to the IMDB scores after the year 2000. You will notice that the dots on the graph are not as clustered after the year 2000 relative to beforehand.

Given this information, I decide that I really want to focus on just the television animated series. Intuitively, I know that a lot of movies produced are relatively low quality straight-to-DVD short films. I can also recall watching some of them as a child and being less than impressed, even back then.

However, I also want to focus on the crossover episodes in particular as well. Intuitively, I think the crossover episodes might have a higher IMDB score, considering it will attract fans of different franchises into the Scooby-Doo universe. I also suspect that crossover episodes might have a higher production budget, but I am unsure.

To recap, I want to filter out anything that isn’t a tv series or crossover episode moving forward to focus my exploratory analysis a little more.



Upon using the view command, I can see a table with all the crossover observations. Interestingly enough, there are only 8 crossover episodes in the dataset.

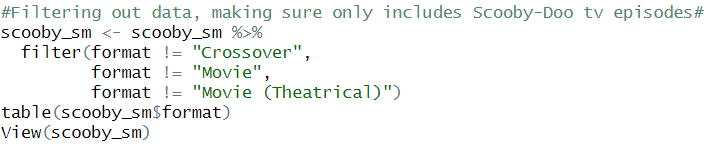


Weird! Scooby-Doo has a crossover with Dynomutt, a series I have personally never heard of, and Supernatural? Scooby-Doo even has a crossover with an Adult Swim adult animation series titled Harvey Birdman. Seems odd. But, interesting! Regardless, just a general question I ask myself from here is why do crossovers even exist, especially with respect to shows I have never even heard of?

Upon further research, I found that crossovers a great way to boost ratings for failing programs. It attracts a larger audience. This is something I had previously intuitively understood, but further research has confirmed my prior suspicions.

For instance, Dynomutt had only lasted 2 seasons with ~20 episodes. And, Harvey Birdman was marginally more successful as an intellectual property, but even it only lasted four seasons ultimately. Inserting Scooby-Doo characters into these franchises via a special crossover episode was essentially a final effort to boost ratings for a dying show.

Because these shows were already failing, I conclude that Scooby-Doo crossovers will likely not affect overall IMDB scores. So, I decide to filter out the crossover episodes as well.



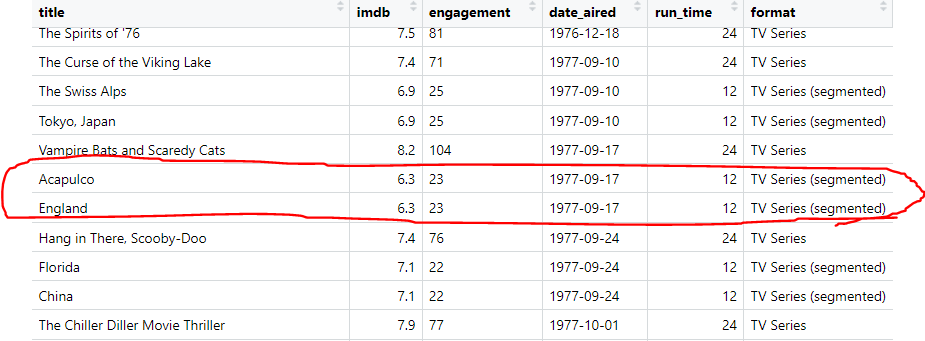
Next, I want to view my newly segmented data frame directly in the table I created.



I can see that my newly segmented dataset now contains only 175 observations. Now, I want to take a peek at these unique 175 episodes in the table.

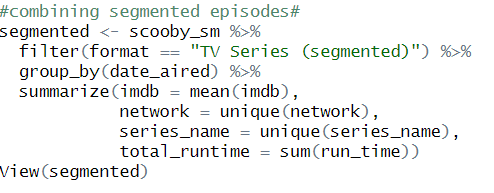


Hmmm, I seem to have encountered a little bit of a potential issue here. There is a general trend I am noticing as I am scrolling through the dataset – segmented TV series are significantly shorter with about ~8-12 minutes of runtime, whereas the data defined by TV series all have a runtime of ~24 minutes. Upon further exploration, I notice these segmented episodes have the same air date too.

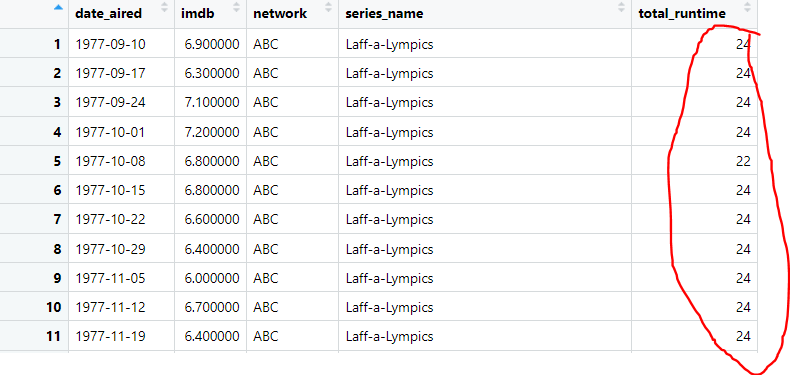


I also notice that the segmented TV series have the same IMDB score in addition. I recognize that this is a problem. If I am to conduct a statistical analysis using this data frame I created, I will essentially be weighting segmented TV series twice as heavily relative to normal TV series episodes. So, my analysis would be fundamentally flawed.

I need to get around this fact. To do so, I need to conduct some data wrangling measures, essentially combining the segmented episodes into a singular episode. That way, it would be an overall fair comparison relative to the TV series category and IMDB scores as a whole.

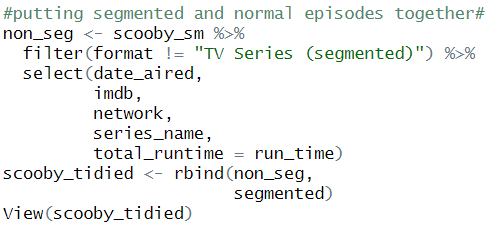


In running this code, I immediately view the table to see if I accomplished what I wanted to accomplish. Likewise, I took the mean score the IMDB episodes just in case some of the IMDB scores differ between the 2 segmented episodes that were aired at the same time.

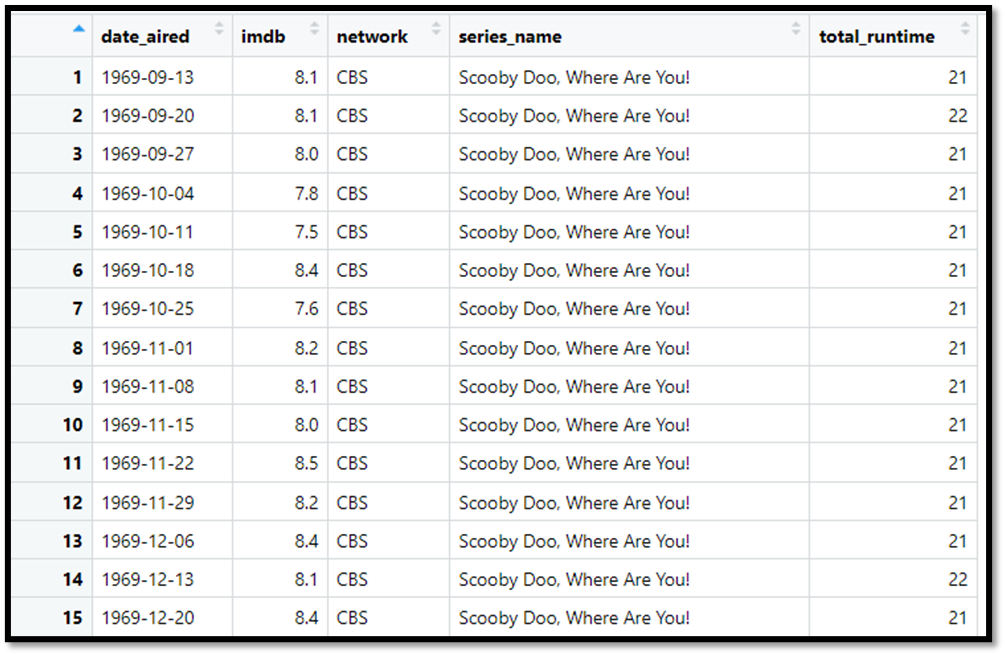


Awesome, I got what I wanted! Now I combined the segmented episodes into a singular aggregate. Essentially, now the segmented episodes are calculated as 1 episode as opposed to 2, and segmented TV episodes as well as normal TV episodes have a similar runtime at ~24 minutes. This avoids that weighting problem I mentioned prior if I were to conduct some statistical analysis.

Now, I need to put that table described above back together with the rest of data. To clarify, I want to put the new aggregate of the segmented episodes back together with the normal TV episodes.

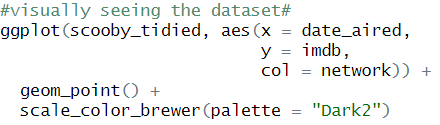


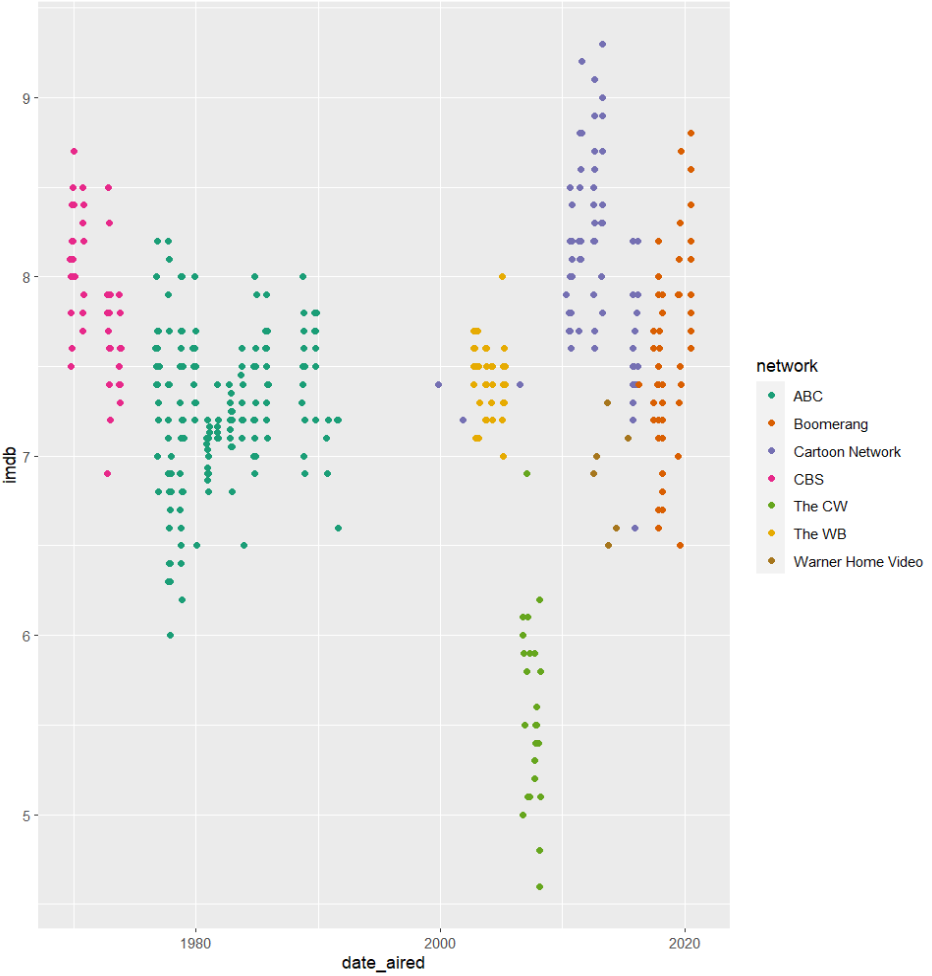
Upon running the code above, I immediately view the newly created table to see if everything worked out.



It appears that everything worked out. I have the columns of the information I want. And, I have a consistent runtime across the board. This newly created table now has 5 columns and 449 rows. This dataset will be the basis for further analysis I will conduct.

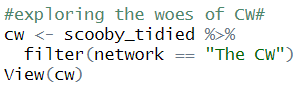
Now, I want visually explore this newly created table of information. Similar to my first example of a scatter plot, I want to see how IMDB scores changed over time. So, I decide that IMDB scores should be on the y-axis, and date aired should be on the x-axis. However, I also want to color and essentially group this information by network. I want to see if the network that published the Scooby-Doo media has an effect on IMDB scores. To do so, I run the following code:



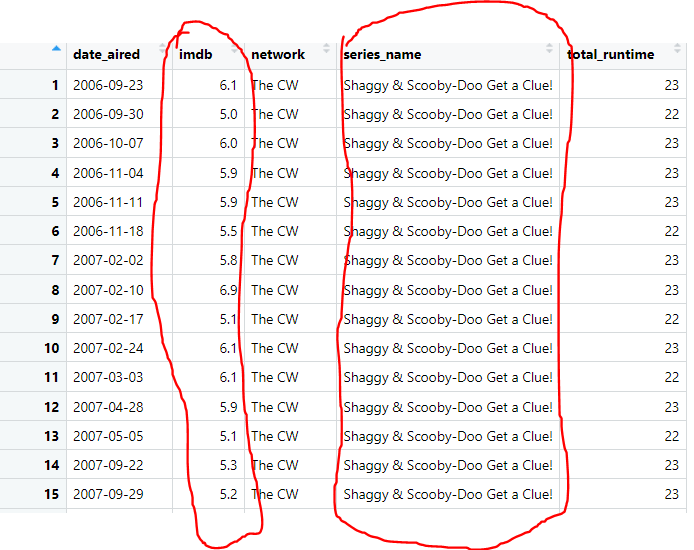
Upon running this code, R produces the following visualization:

This is the visualization that I was looking for! From the illustration, you see IMDB scores started out strong under CBS’s leadership, then declined over time under ABC’s leadership. Furthermore, IMDB ratings hit an all-time low under CW’s leadership, but the franchise bounced back reaching record highs under Cartoon Network and Boomerang leadership respectively.

As I previously noted, the CW really dropped the ball when it comes to Scooby-Doo franchise. What the heck was going on there? I want to explore this more deeply. To do so, I run the following code:



Upon running the code, I want to view the newly created table. This table will contain all Scooby-Doo related information published under the CW network umbrella.



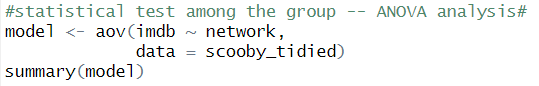
We can see from the table that IMDB scores were frankly extremely poor. Likewise, upon further exploration, I notice that the series name published under CW was Shaggy & Scooby-Doo Get a Clue! Clearly, fans of the Scooby-Doo franchise were not receptive to this series in the slightest. Let me dig a little bit deeper.

After doing a bit of searching online, I found that Shaggy & Scooby-Doo Get a Clue! was different relative to other shows in the Scooby-Doo franchise in a number of ways. As an illustration, fans of Scooby-Doo were disappointed by the different animation style. Likewise, many members of the Scooby-Doo gang are not present at all during this series. For instance, Velma, Fred, and Daphne rarely make an appearance.

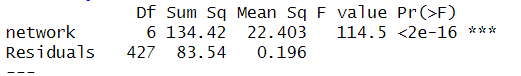
To recap, I already know this showed failed. But, upon further research, I believe Shaggy & Scooby-Doo failed because it strayed to far from the norm. Scooby-Doo is a large franchise that had built a successful formula over the years. And, I think the audience just did not take kindly to the types of changes that were made.

Referring back to the original visualization produced, we can see that there are differences in IMDB scores among networks. However, I want to confirm this assumption by running some statistical tests. I want to confirm if these differences are statistically significant.

To do so, I decide to conduct a test that would report an analysis of variance. This is commonly considered as an ANOVA analysis. Variance, in this case, would be differences among the networks. In particular, I want to see if the IMDB scores under CW’s leadership is significantly different among the other networks. I won’t necessarily delve into the specifics of this statistical test; however, you can think of it as an extension of a typical t-test for independent samples but for more than two groups. To run the ANOVA analysis, I produce the following code:

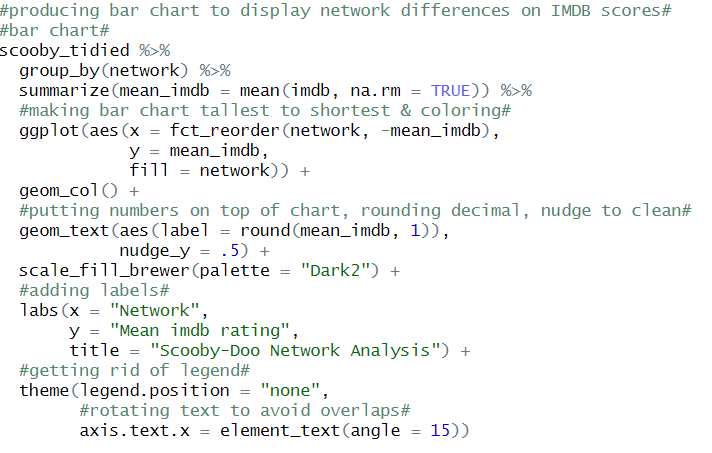


From the test, we can see the following:

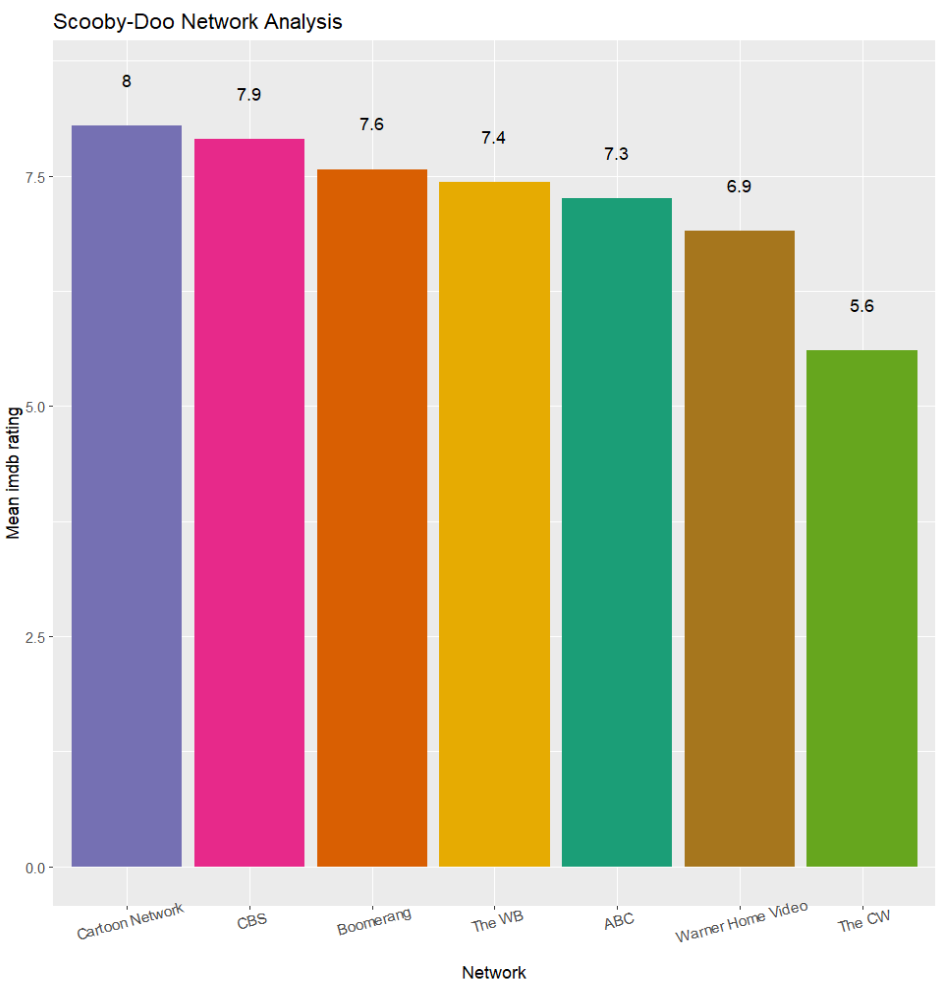


From the table provided above, we see that the p-value is very small, and it is extremely statistically significant. This test pretty much just confirms the CW’s poor IMDB scores are extraordinarily likely to be different than the other networks – it was not just a coincidence that the Scooby-Doo franchise had poor IMDB scores under the CW’s leadership.

Finally, I just want to produce a bar chart to display average IMDB ratings among the different networks. In doing so, I can get a better understanding of the differences between the networks and how their influence effected the Scooby-Doo franchise. To do so, I run the following code in R:



In running this code, the end product is the following visualization:

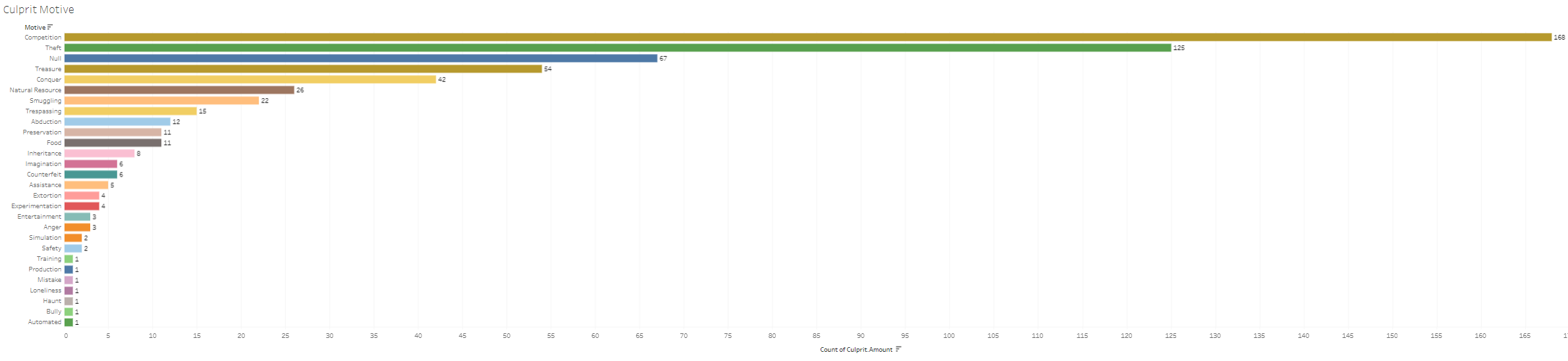


From the visual, we can clearly identify that IMDB ratings were the highest among Cartoon Network’s leadership at 8/10. Likewise, we can that the Scooby-Doo franchise declined in quality under CW’s leadership to 5.6/10. Also, we can identify, in order, which networks had the best ratings in relation to Scooby-Doo media properties – Cartoon Network, CBS, Boomerang, Warner Brothers, ABC, Warner Home Video, and CW.

To reiterate, the purpose of this analysis was to demonstrate my proficiency using the R programming language. I also demonstrated my ability to clean and wrangle data effectively, conduct exploratory data analysis, utilize relevant statistical tests, and produce visualizations that help tell the story of the data. While this is a project sort of for fun and practice, many of the same methods can be applied and are directly transferrable to a more professional setting.

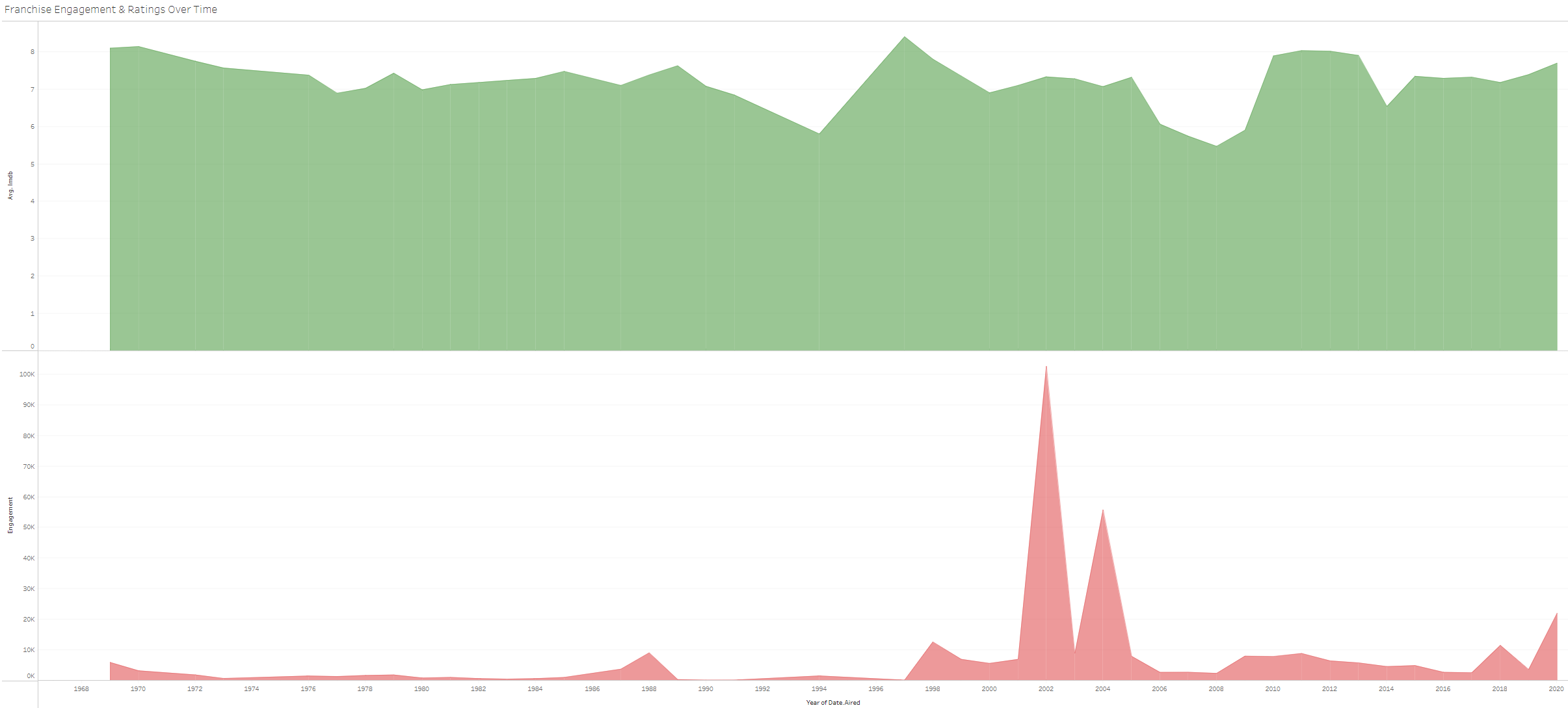
Now, I want to transition towards working with Tableau to further explore the Scooby-Doo universe. Please, access this link, which was aforementioned at the very beginning of this paper, to follow along: <https://public.tableau.com/views/VisualExplorationofScooby-DooUniverse/CulpritMotive?:language=en-US&:display_count=n&:origin=viz_share_link>.

Firstly, I am curious to see what the underlying motives are for culprits committing crimes in the Scooby-Doo universe.



I understand the image may be difficult to see. Please, follow the link provided above to see the visualization in a better format. Regardless, from the visual, I can clearly identify the motives of criminals in the Scooby-Doo universe. As an illustration, the primary motives for most villains seems to be competition, theft, and treasure – in that order. Likewise, surprisingly, one villain was classified to have the motive of a simple mistake too, which I found humorous. Moreover, 3 truly evil individuals seemed to commit crimes simply due to an entertainment factor.

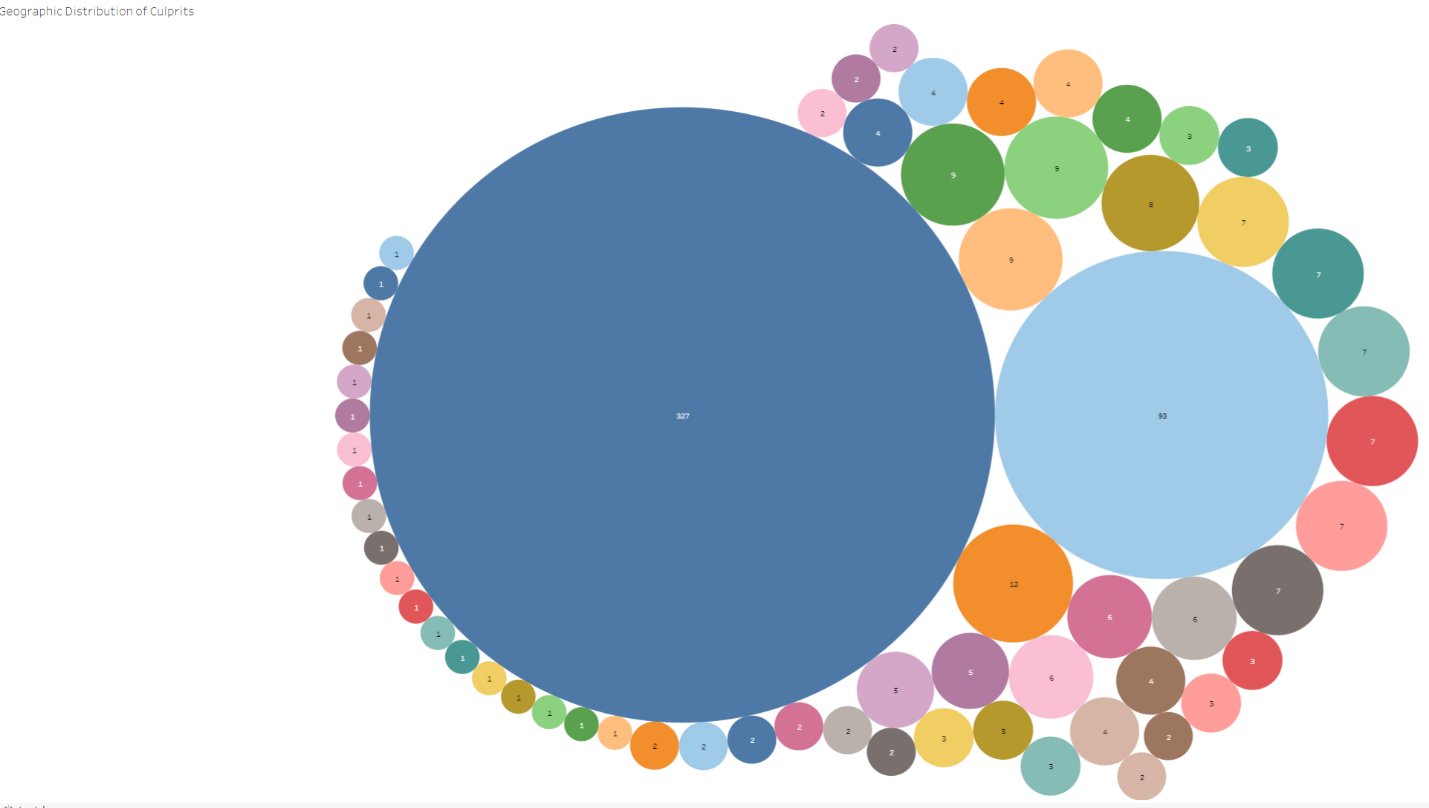
Next, I wanted to explore the correlation between franchise engagement and ratings over time. To do so, I produced two separate area charts displayed on the same page.



The green area chart is average IMDB scores over time, and the red area chart is engagement over time. You can think of average IMDB scores as a proxy for quality. To my surprise, it appears quality and engagement are not correlated whatsoever. If they were, we expect the chart to have similar characteristics and move in similar directions.

In addition, it appears that engagement in the Scooby-Doo franchise hit an all-time high in 2002, and average IMDB scores hit an all-time in 1997. This is yet another indication that engagement and quality are not correlated. If they were, we expect highs for both charts to occur in the same year.

Next, I wanted to explore the geographic distribution of culprits in the Scooby-Doo universe. Initially, I was envisioning a map to explore this relationship; however, the dataset had a few issues. For instance, sometimes geographic origin is defined as United States, but other times geographic origin is defined by a state. The same problem occurs for other countries as well. So, ultimately, I decided that the most appropriate chart to use for this analysis would be a bubble chart.





Naturally, you can hover over each individual circle to see exact details in Tableau. Nevertheless, it appears most culprits originate from the United States, which makes sense considering Scooby-Doo as an intellectual property has its origin in the United States. Similarly, a vast majority of these criminals actually originate from California in the United States, which is represented by the large light blue circle. Each individual circle’s size is proportionate to the number of criminals. And, each individual circle represents a unique geographic location with respect to a culprit’s origin.

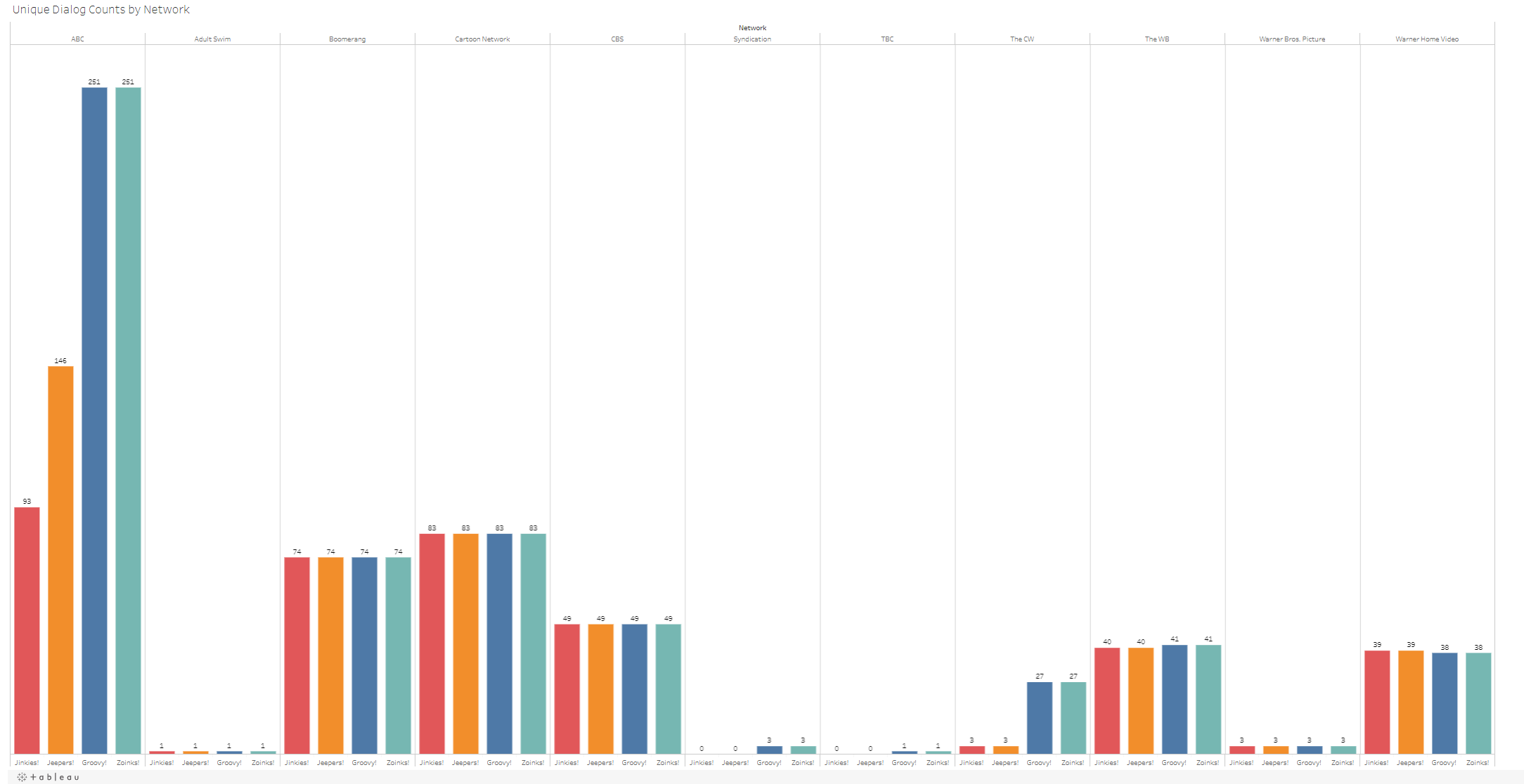
Something that I found interesting here is that there was only 1 occurrence of a culprit’s origin in Russia. I found this interesting considering Scooby-Doo has been around a long time, and the franchise was popular and active during Cold War. Normally, I would expect American media to be a reflection of the poor relationship among the superpowers. But I guess Scooby-Doo might have taken the high road? Truthfully, I am unsure; nevertheless, I found it rather interesting.

Next, I wanted to explore unique dialogue occurrences throughout Scooby-Doo’s history. If you are not familiar with the franchise, certain characters have unique dialogue.



Pictured above is the main cast of the Scooby-Doo franchise. From left to right, we have Fred, Velma, Scooby himself, Shaggy, and Daphne. To relate back to what I was saying prior, certain characters have unique dialogue. As an illustration, Shaggy often says “zoinks!,” Velma often says “jinkies!,” Daphne says “jeepers!,” etc.

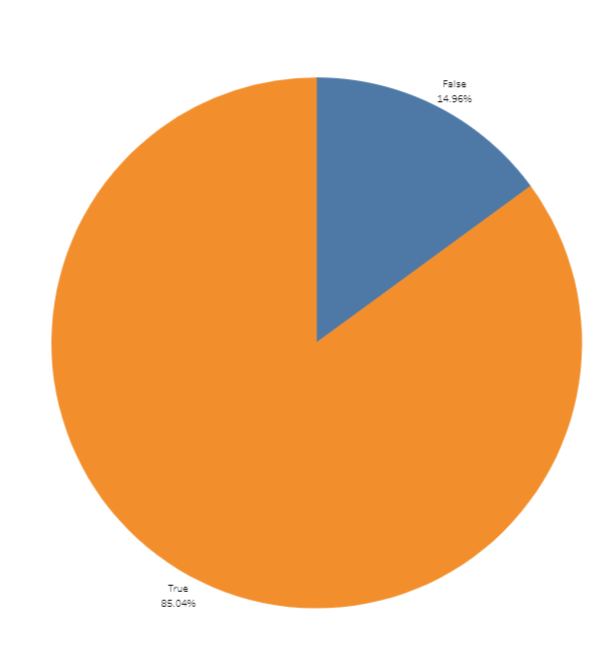
I wanted to explore if there were any differences among the networks relative to the number of times a character’s unique catchphrase is exclaimed. To explore this relationship, I produced the following visualization.



As you can see, there were differences among the networks, as suspected. The red bar charts represent “jinkies!,” the orange represent “jeepers!,” the dark blue represent “groovy!,” and the light blue represent “zoinks!.”

From the illustration, we can see that ABC used every category of catchphrase the most relative to other networks. In fact, ABC was particularly a fan of “groovy!” and “zoinks!” apparently. Similarly, there is variation in the occurrences of these catchphrases with respect to ABC as well. To my surprise, other networks seemingly used the catchphrases much more conservatively, and there is little variation. Perhaps, the different networks had a quota of using each catchphrase once or twice during an episode. For instance, both Cartoon Network used the catchphrases the same number of times across the board – at 74 times and 83 times respectively.

Next, I wanted to see the rate at which captured criminals are actually arrested in the franchise. Intuitively, I figured that every criminal would be arrested; however, to my surprise, this was not the case.

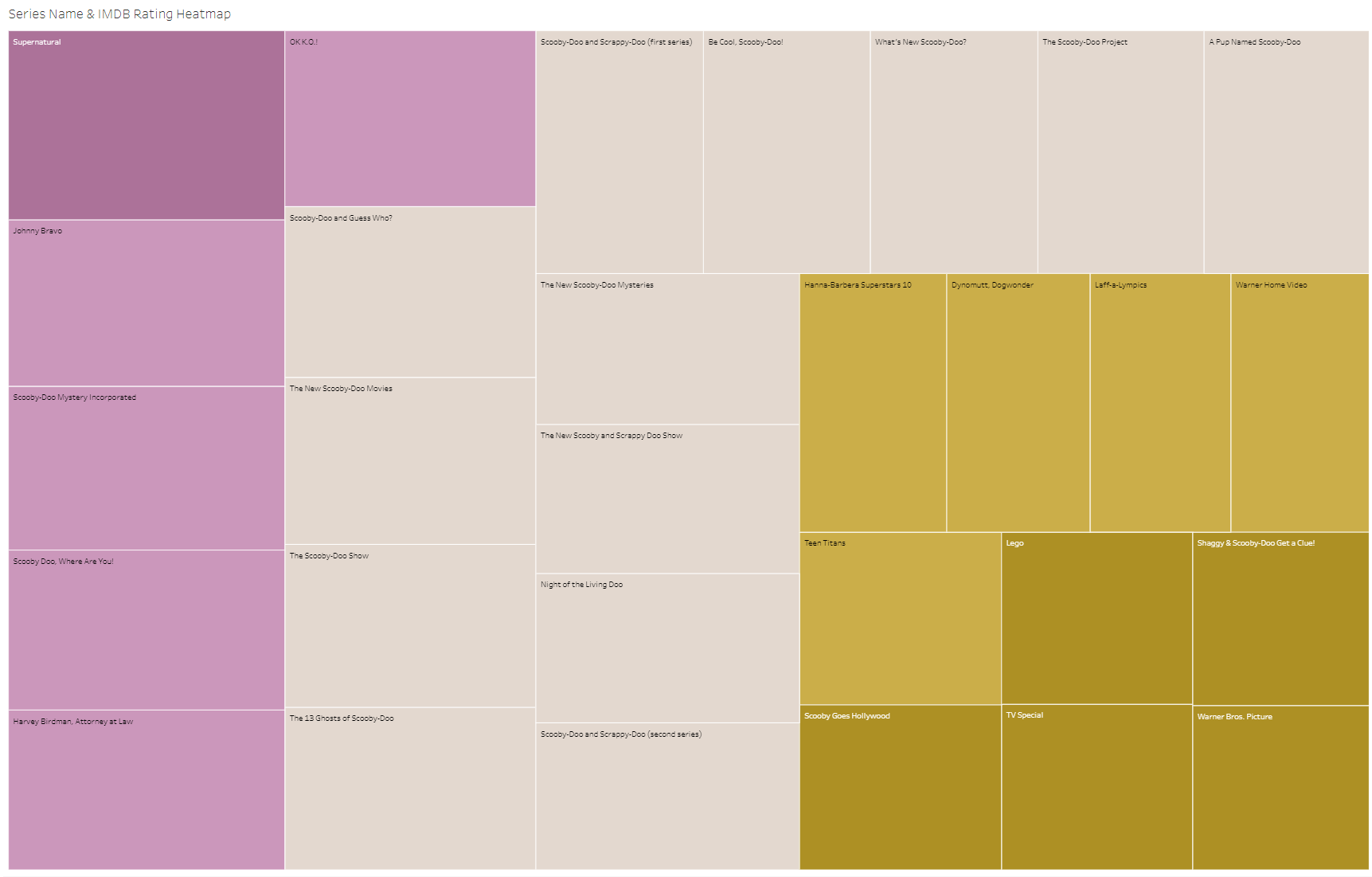


I categorized if a culprit was arrested as a true/false statement in the dataset. From the visual, the orange part of the pie represents criminals who were, indeed, arrested. This constituted 381 convicted criminals. And, the Scooby-Doo gang have a conviction rate of ~85%. WOW!

Upon further research, I found that the felony conviction rate among criminals tried by the state or county is about ~95%. These trials are spearheaded by some the best lawyers in the world. Yet, the Scooby-Doo gang is just a ragtag group of free-spirited individuals. For them to have a conviction rate on par with federal attorneys is quite remarkable to say the least!

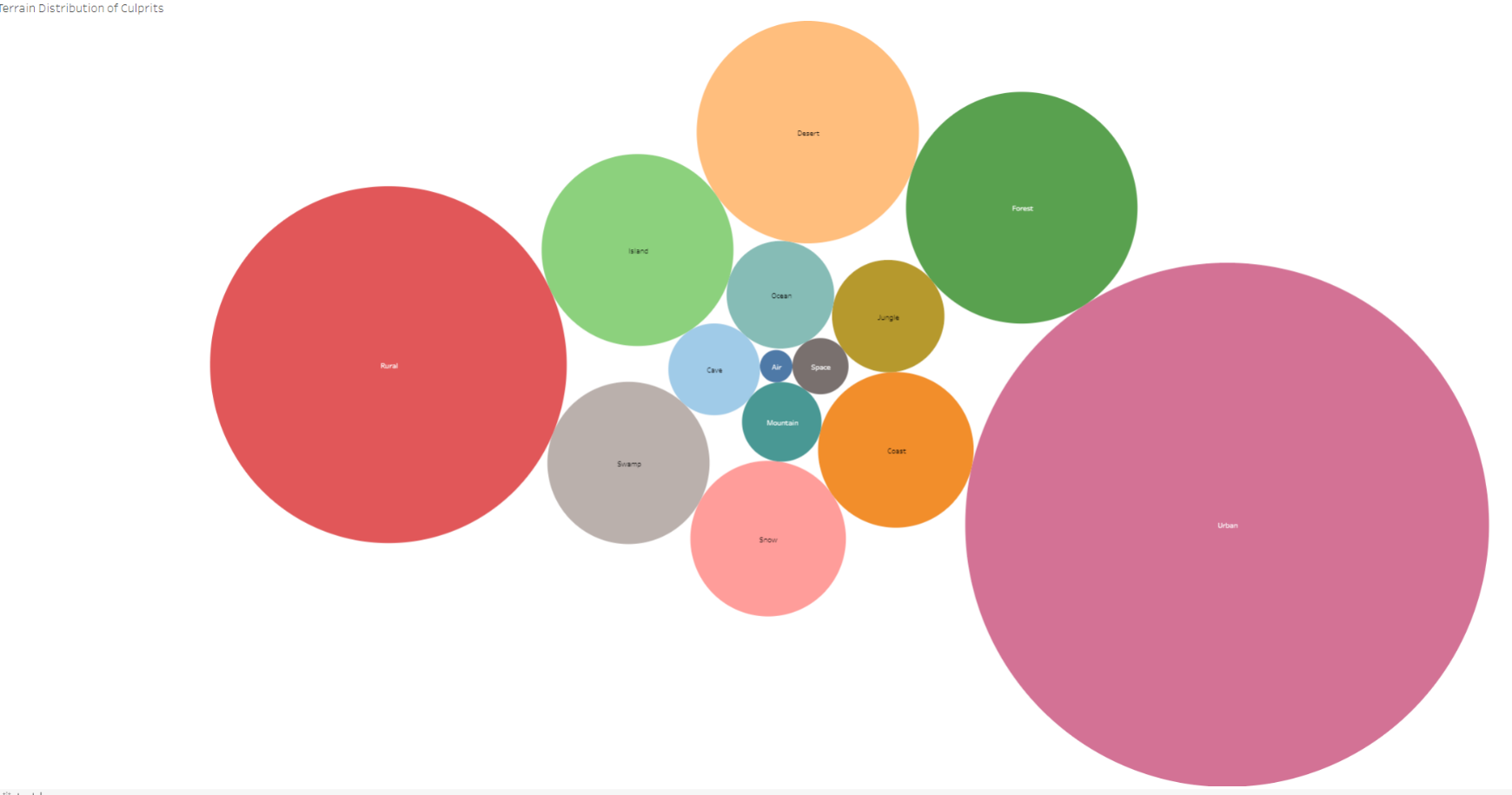
Conversely, we can see from the visualization the blue slice as well. The blue slice represents criminals who were not arrested after the Scooby-Doo gang captured them. This slice constitutes 67 criminals and the remaining 15% of the pie. As a whole, it has been proven through this visual analysis that the gang is pretty damn good, just not perfect.

Next, I wanted to explore the relationship between the series name within the Scooby-Doo franchise and IMDB scores. Ultimately, I produced a heatmap to conduct my visual analysis.



From this visualization, we can identify that the crossover episode with Supernatural, located in the top right of the diagram, is actually, historically, the highest-rated episode within the Scooby-Doo universe. We can also see that a movie produced by Warner Brothers Pictures is, historically, the lowest rated. As expected, you can hover over each individual square in Tableau to see further details.

Now, I wanted to revisit culprit origin. As aforementioned, the data was not necessarily the best with respect to geographic information. However, I do have data with respect to the terrain in which each episode took place. I can use this as a proxy to define culprit origin.



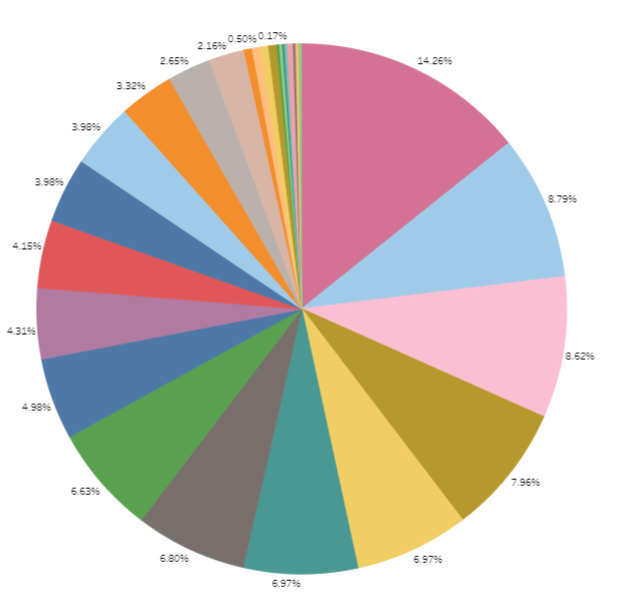
The Scooby-Doo

As an illustration, Shaggy, pictured below,

Firstly, I am curious to see what the underlying motives are for culprits committing crimes in the Scooby-Doo universe.

Using this information, I can clearly see that most culprits commit crimes in urban areas. This is represented by the largest purple circle on the right. Similarly, the next most common terrain type as rural, represented by the red circle. To my surprise, there are some culprits the Scooby-Doo gang tracked down in space! This is represented by the small dark grey square in the center. Like the other diagrams, you can hover over each circle for more details.

Lastly, I wanted to see the culprit distribution by series. To do so, I produced a typical pie chart. You can hover over each slice for further details naturally.



From the diagram, we can see that Scooby-Doo & Scrappy-Doo is the series that contained the most culprits. This is represented by the largest slice of pie, pictured in a deep pink color. In fact, Scooby-Doo & Scrappy-Doo constituted about ~14% of all the criminals throughout the historic Scooby-Doo franchise, and this amounted to exactly 86 criminals. Scooby-Doo & Scrappy-Doo is followed by Be Cool, Scooby-Doo! and Scooby-Doo Mystery Incorporated respectively. In fact, these three series alone accounted for almost ~25% of all criminals throughout the multi-decade franchise!

The purpose of this exercise was to take my findings of R a step further, utilizing Tableau for advanced data exploration and visual analysis. I demonstrated my proficiency working with Tableau, and my ability to produce appealing and informative illustrations that tell the story of the data. Like my analysis with R, this analysis was mainly for practice and fun; however, the exact same concepts can be directly applied and are directly transferrable to a more professional setting. Again, please access the following link to interact with the visualizations produced in a better format: <https://public.tableau.com/views/VisualExplorationofScooby-DooUniverse/CulpritMotive?:language=en-US&:display_count=n&:origin=viz_share_link>.

Thanks for Reading!!!!