# ADV Stats: Final Project

## ## Introduction: The Road So Far…

*Supernatural* was a long-running TV series that became a cultural phenomenon, airing on The CW for an impressive 15 seasons and 16 years, from 2005 to 2021. The show followed the Winchester brothers, Sam and Dean, as they traveled across America battling supernatural creatures, solving mysteries, and ultimately throwing punches at God himself. What started as a series inspired by urban legends and ghost stories evolved into an epic saga filled with action, humor, and emotional depth. By the time the finale aired, *Supernatural* had delivered 327 episodes packed with intricate lore, memorable characters, and poignant moments.

While the series concluded on a bittersweet note, trying to tie up many as many of its loose ends as possible, the reception to it's finale was mixed at best. Fans, still passionate about the show, often ask the actors behind Sam and Dean, Jared Padalecki and Jensen Ackles, whether they would ever reprise their roles. Their coy response of "Ask us in 2025" has only fueled speculation about the possibility of reviving the series.

Now, with 2025 just around the corner, it feels like the perfect time to imagine what a new season might look like; after all, if I've learned anything from this show, it's that nothing stays dead for long. But after such a long run, the later seasons of *Supernatural* struggled to maintain the same level of critical and audience acclaim as earlier episodes. To ensure a hypothetical Season 16 wouldn't face the same challenges, I've conducted an Exploratory Data Analysis (EDA) of the show's first 15 seasons. My goal? To analyze what resonated most with viewers and create a plan to maximize Season 16's chances of success.

## ## Step 1: Data Collection

The data for this analysis was sourced from IMDb - The Internet Movie Database

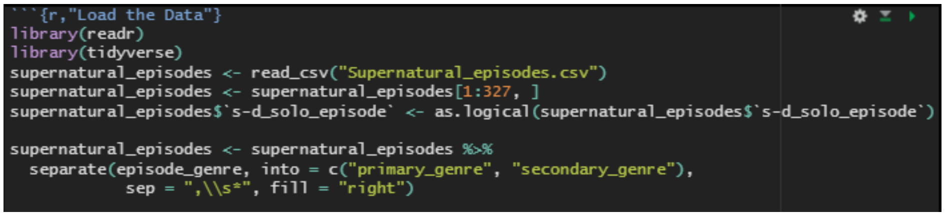
Observations: 327

Variables: 15

- episode_id: The overall number of the episode (1–327).
- episode_position: The episode's position within its season (1–16/20/22/23).
- season_number: The season to which the episode belongs.
- episode_title: The title of the episode.
- director: The name of the episode's director.
- writer: The writer(s) of the episode.
- air_date: The date the episode aired on The CW.
- days_between_episodes: How many days passed between episodes (*aggregated variable*).
- views_at_air_mil: How many people (in millions) watched the episode on its air date.
- imdb_rating: The IMDb rating of the episode.
- mow_filler: Whether the episode was a Monster-of-the-Week-style episode or filler (*Boolean logic*).
- episode_genre: The episode's primary genre and subgenre (when applicable).
- centric_characters: Which main characters were central to the episode.
- s_d_solo_episode: Whether the episode was a Sam-and-Dean solo episode (*Boolean logic; aggregated variable*).
- episode_description: The episode's description as listed on IMDb.

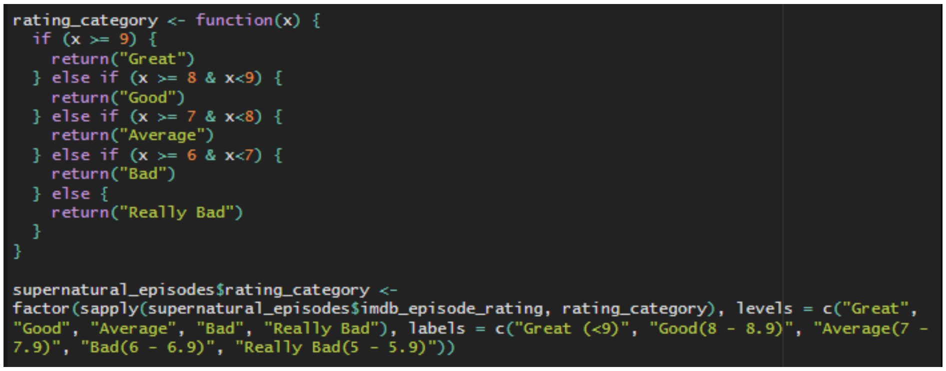
The primary variables used in this analysis include: imdb_rating, episode_genre, centric_characters, mow_filler, season_number, and episode_id

Before I began my analysis, I needed to perform some general data cleaning and a couple of data aggregations. The code below loaded my dataset into RStudio, removed the NA columns, and separated the episode_genre category into Primary and Secondary Genres.



~~~
```{r,"Load the Data"}
library(readr)
library(tidyverse)
supernatural_episodes <- read_csv("Supernatural_episodes.csv")
supernatural_episodes <- supernatural_episodes[1:327, ]
supernatural_episodes$`s-d_solo_episode` <- as.logical(supernatural_episodes$`s-d_solo_episode`)

supernatural_episodes <- supernatural_episodes %>%
  separate(episode_genre, into = c("primary_genre", "secondary_genre"),
           sep = ",\\s*", fill = "right")
~~~

Meanwhile, this code created a new column called rating_category, which allowed me to classify all of the imdb_episode_ratings into categorical factors, which I used to color my data visualizations.



```
rating_category <- function(x) {
  if (x >= 9) {
    return("Great")
  } else if (x >= 8 & x<9) {
    return("Good")
  } else if (x >= 7 & x<8) {
    return("Average")
  } else if (x >= 6 & x<7) {
    return("Bad")
  } else {
    return("Really Bad")
  }
}

supernatural_episodes$rating_category <-
factor(sapply(supernatural_episodes$imdb_episode_rating, rating_category), levels = c("Great",
"Good", "Average", "Bad", "Really Bad"), labels = c("Great (<9)", "Good(8 - 8.9)", "Average(7 -
7.9)", "Bad(6 - 6.9)", "Really Bad(5 - 5.9)"))
```

## ## Step 2: Written Documentation

**Problem Description:** The goal of this analysis is to explore what worked well in the first 15 seasons of Supernatural and use those insights to guide the development of Supernatural: Season 16. By examining key patterns from the show's past, I aim to provide recommendations on several crucial aspects for the new season, based on the following research questions:

1. *Primary Genre Focus:* Which genre most resonated with audiences across plot-driven episodes in the first 15 seasons? I plan on using this as the primary genre for the Season 16 plot
2. *Genre Combinations:* Which primary and secondary genre combinations led to the highest average episode ratings?
3. *MotW/Filler Ratio:* What is the ideal balance between Monster-of-the-Week (MotW) (episodes with self-contained plots and conclusions) episodes and plot-driven episodes? Is there a correlation between MotW percentage and overall Season reception?
4. *Characters:* What character(s) had the highest episode ratings throughout the series?

**Related Work:** This analysis was primarily inspired by the TV showing graphs that get distributed once a series finale airs. This analysis was specifically inspired by this post:

Reddit - Dive into anything

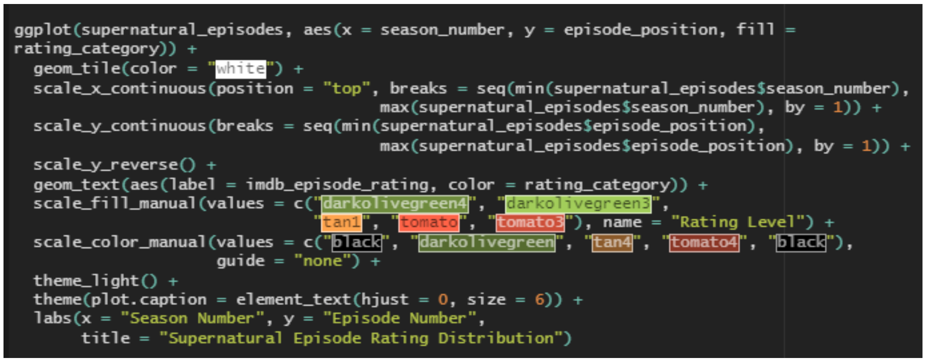
where a Reddit user shared an episode rating distribution visualization, which I sought to recreate in my preliminary EDA process (visualization 1). However, I personally haven't seen a genre analysis like this done in any official capacity, so my curiosity drove me to do it myself.

**Solution:** This analysis primarily used visual distribution (Module 7) and correlation techniques (Module 8), alongside basic time series analysis (module 10) and visual multivariates (Module 9). Initial exploration includes visualizing the distribution of episode ratings across seasons, while heatmaps and box plots are used to identify the most effective genre combinations for Season 16. Additionally, I explored the optimal percentage of MotW episodes for the new season by conducting correlation and time series analysis, examining how the average ratings of MotW and plot-driven episodes evolved over the show's run (episodes 1-327, 2005-2021), as well as how the percentage of MotW episodes correlates with season ratings. To determine which characters should take center stage, I analyze their average episode ratings using bar charts and visual distribution analysis.

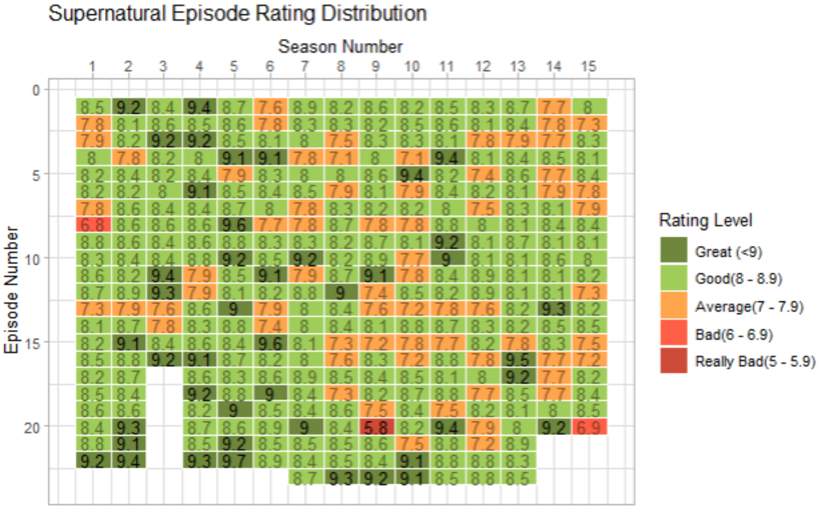
## ## Step 3: Data Visualizations

Visualization 1: EDA

This first visualization was created primarily for data exploration. The goal was to better understand the dataset's structure and examine how IMDb ratings were distributed across the 15 seasons of *Supernatural*. I chose to use a heatmap for this purpose, with the X-axis representing the season number, the Y-axis showing the episode position within each season, and the color of each tile indicating the episode's rating category, as defined during the data preprocessing (or "data munging") stage.



```
ggplot(supernatural_episodes, aes(x = season_number, y = episode_position, fill =
rating_category)) +
  geom_tile(color = "white") +
  scale_x_continuous(position = "top", breaks = seq(min(supernatural_episodes$season_number),
                                                    max(supernatural_episodes$season_number), by = 1)) +
  scale_y_continuous(breaks = seq(min(supernatural_episodes$episode_position),
                                                    max(supernatural_episodes$episode_position), by = 1)) +
  scale_y_reverse() +
  geom_text(aes(label = imdb_episode_rating, color = rating_category)) +
  scale_fill_manual(values = c("darkolivegreen4", "darkolivegreen3",
                                "tan1", "tomato", "tomato3"), name = "Rating Level") +
  scale_color_manual(values = c("black", "darkolivegreen", "tan4", "tomato4", "black"),
                    guide = "none") +
  theme_light() +
  theme(plot.caption = element_text(hjust = 0, size = 6)) +
  labs(x = "Season Number", y = "Episode Number",
       title = "Supernatural Episode Rating Distribution")
```

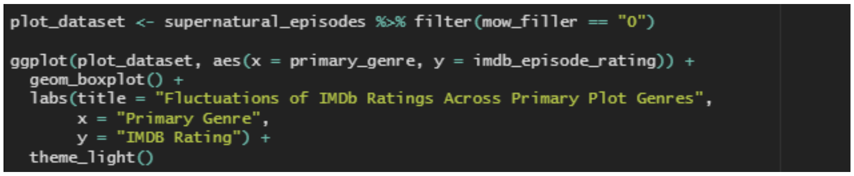


This chart demonstrates that, while ratings for *Supernatural* gradually declined as the series progressed, they remained relatively stable throughout its 15-season run. This suggests that there is still strong potential for success with a hypothetical Season 16.

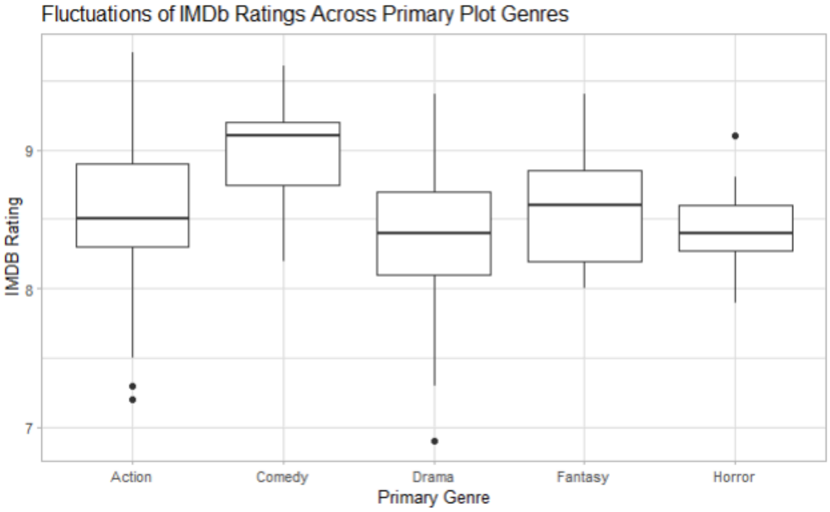
Visualization 2: Primary Genre's

With a clearer understanding of my dataset, I moved on to address the first question I posed for designing a hypothetical Season 16: Which genre should the overarching plot follow?

To tackle this, I filtered the dataset to include only plot-related episodes and created a box plot. This visualization highlights the differences in mean and quantile distributions across the primary genre categories, providing insight into which genres have historically performed better in terms of IMDb ratings.



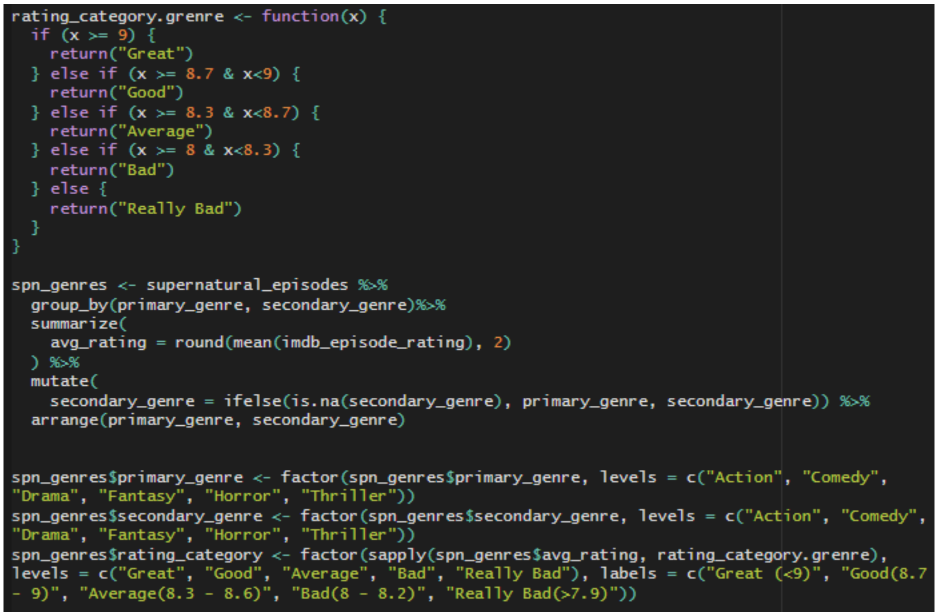
```
plot_dataset <- supernatural_episodes %>% filter(mow_filler == "0")

ggplot(plot_dataset, aes(x = primary_genre, y = imdb_episode_rating)) +
  geom_boxplot() +
  labs(title = "Fluctuations of IMDb Ratings Across Primary Plot Genres",
       x = "Primary Genre",
       y = "IMDB Rating") +
  theme_light()
```



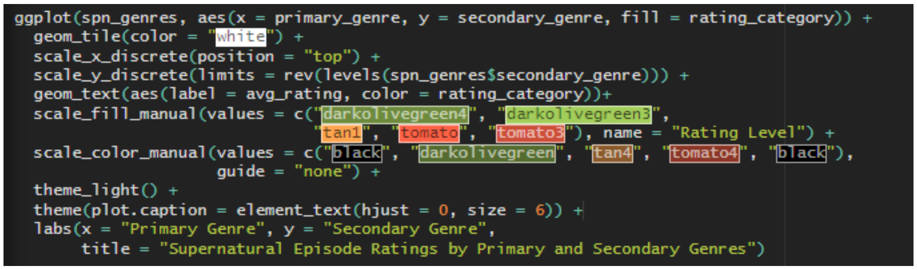
This visualization reveals that plot-driven episodes with a primarily comedic focus outperformed all other genre categories. Notably, comedy was the only primary genre to achieve an average IMDb rating higher than 9.0, making it a standout choice for engaging audiences.

Visualization 3: Genre Combinations

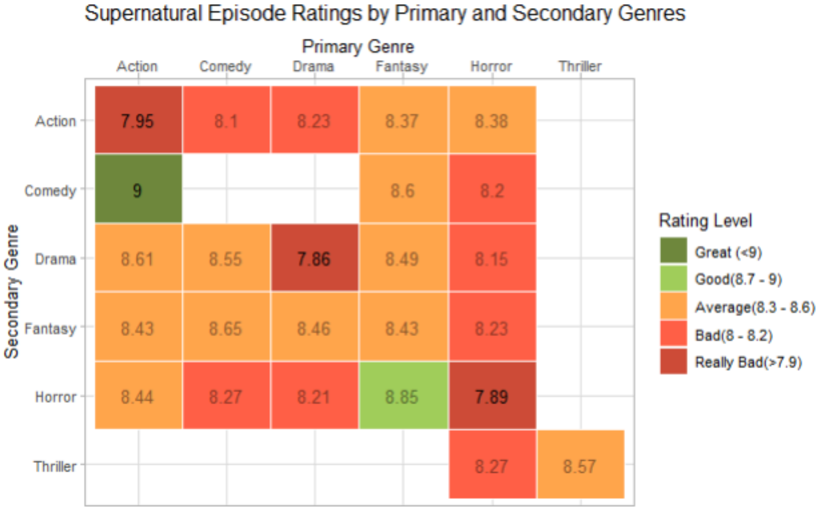
The code for analyzing how genres interact with each other was a bit more complex than the previous examples. To determine which genre combinations were most effective, I grouped the dataset by both primary and secondary genres and then calculated the average IMDb rating for each combination using dplyr's summarize() function. Once I had the grouped data, I converted the genres into categorical factors to ensure proper ordering and readability. Finally, I visualized the results using a heatmap, applying a similar approach to the code used for Visualization 1. This allowed me to effectively display how different genre pairings influenced audience ratings.



```
rating_category.grenre <- function(x) {
  if (x >= 9) {
    return("Great")
  } else if (x >= 8.7 & x<9) {
    return("Good")
  } else if (x >= 8.3 & x<8.7) {
    return("Average")
  } else if (x >= 8 & x<8.3) {
    return("Bad")
  } else {
    return("Really Bad")
  }
}

spn_genres <- supernatural_episodes %>%
  group_by(primary_genre, secondary_genre)%>%
  summarize(
    avg_rating = round(mean(imdb_episode_rating), 2)
  ) %>%
  mutate(
    secondary_genre = ifelse(is.na(secondary_genre), primary_genre, secondary_genre)) %>%
  arrange(primary_genre, secondary_genre)


spn_genres$primary_genre <- factor(spn_genres$primary_genre, levels = c("Action", "Comedy",
"Drama", "Fantasy", "Horror", "Thriller"))
spn_genres$secondary_genre <- factor(spn_genres$secondary_genre, levels = c("Action", "Comedy",
"Drama", "Fantasy", "Horror", "Thriller"))
spn_genres$rating_category <- factor(sapply(spn_genres$avg_rating, rating_category.grenre),
levels = c("Great", "Good", "Average", "Bad", "Really Bad"), labels = c("Great (<9)", "Good(8.7
- 9)", "Average(8.3 - 8.6)", "Bad(8 - 8.2)", "Really Bad(>7.9)"))
```



```
ggplot(spn_genres, aes(x = primary_genre, y = secondary_genre, fill = rating_category)) +
  geom_tile(color = "white") +
  scale_x_discrete(position = "top") +
  scale_y_discrete(limits = rev(levels(spn_genres$secondary_genre))) +
  geom_text(aes(label = avg_rating, color = rating_category))+
  scale_fill_manual(values = c("darkolivegreen4", "darkolivegreen3",
                               "tan1", "tomato", "tomato3"), name = "Rating Level") +
  scale_color_manual(values = c("black", "darkolivegreen", "tan4", "tomato4", "black"),
                     guide = "none") +
  theme_light() +
  theme(plot.caption = element_text(hjust = 0, size = 6)) +
  labs(x = "Primary Genre", y = "Secondary Genre",
       title = "Supernatural Episode Ratings by Primary and Secondary Genres")
```



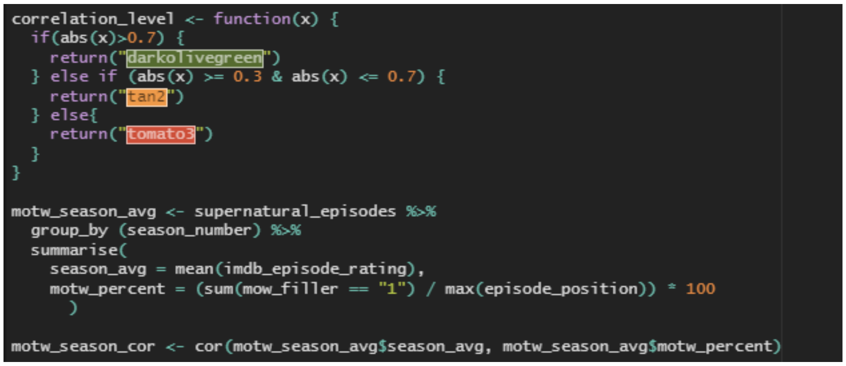
What I found was that single-genre episodes were by far the worst-performing category, whereas Action-Comedies and Fantasy-Horrors stood out as the highest-rated episode combinations. This bodes well for our Season 16 design, as pairing a comedic overarching plot with action-comedy episodes aligns with these findings.

One surprising insight from this chart was that horror episodes, despite the show's origins as a horror-based Monster-of-the-Week series, performed poorly overall. This may be attributed to the challenges of executing horror effectively in a television format with limited budgets, where poorly executed horror can come across as cheesy rather than compelling.

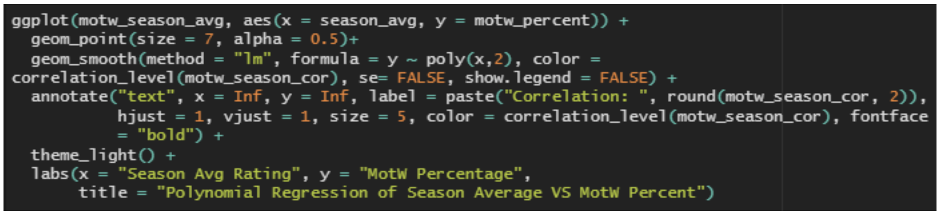
Fantasy elements in the supernatural usually pertain to different mythologies, such as Christian allegories, Norse gods, and Greek monsters.

Visualization 4/5: MotW / Filler Ratio

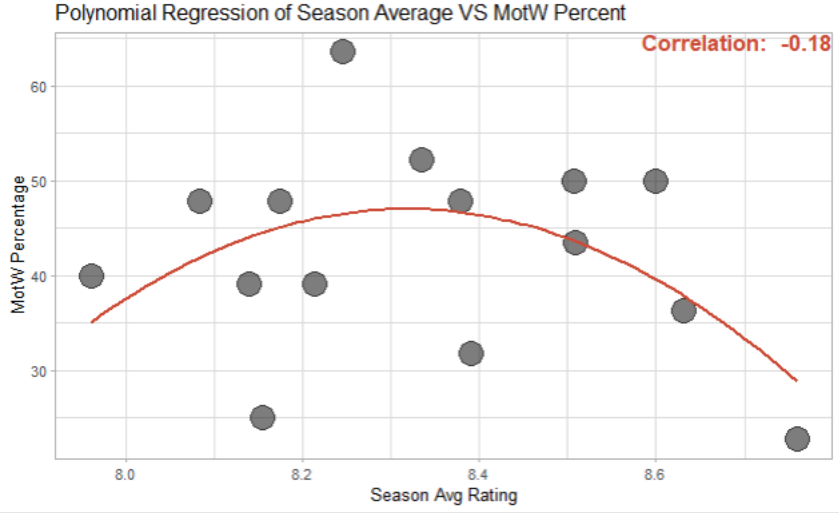
With the primary genre of Season 16 decided and the types of episodes identified, the next step was to determine what percentage of filler episodes should make up the season. To address this, I created a scatter plot to analyze the relationship between the percentage of MotW filler episodes each season and the overall season IMDb rating. To better capture the trend, I added a polynomial line of best fit, which would have highlighted the correlation between filler content and season performance. Unfortunately, as you will see, there does not appear to be any significant correlation between the two.



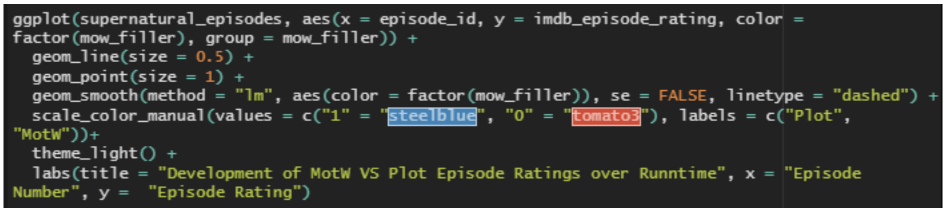
```
correlation_level <- function(x) {
  if(abs(x)>0.7) {
    return("darkolivegreen")
  } else if (abs(x) >= 0.3 & abs(x) <= 0.7) {
    return("tan2")
  } else{
    return("tomato3")
  }
}

motw_season_avg <- supernatural_episodes %>%
  group_by (season_number) %>%
  summarise(
    season_avg = mean(imdb_episode_rating),
    motw_percent = (sum(mow_filler == "1") / max(episode_position)) * 100
      )

motw_season_cor <- cor(motw_season_avg$season_avg, motw_season_avg$motw_percent)
```



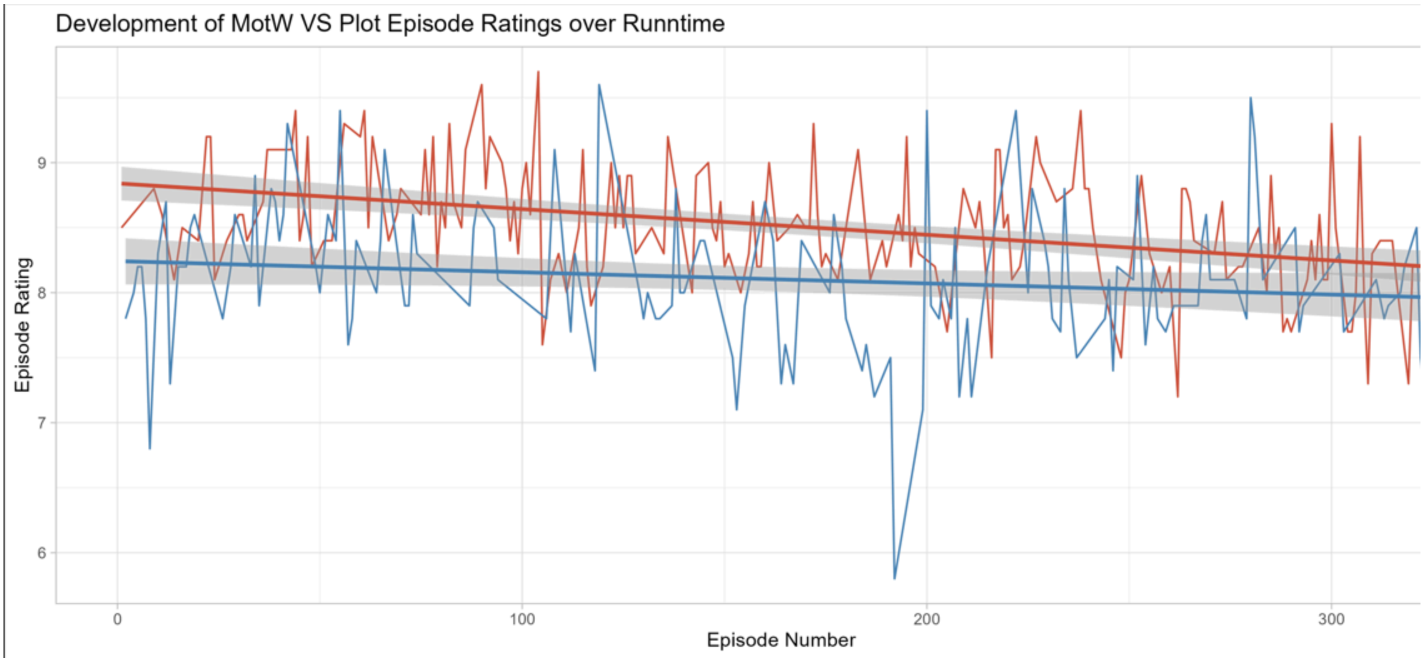
```
ggplot(motw_season_avg, aes(x = season_avg, y = motw_percent)) +
  geom_point(size = 7, alpha = 0.5)+
  geom_smooth(method = "lm", formula = y ~ poly(x,2), color =
correlation_level(motw_season_cor), se= FALSE, show.legend = FALSE) +
  annotate("text", x = Inf, y = Inf, label = paste("Correlation: ", round(motw_season_cor, 2)),
           hjust = 1, vjust = 1, size = 5, color = correlation_level(motw_season_cor), fontface
           = "bold") +
  theme_light() +
  labs(x = "Season Avg Rating", y = "MotW Percentage",
       title = "Polynomial Regression of Season Average VS MotW Percent")
```



With no apparent correlation between these two factors, I shifted my focus to examining the development of MotW episodes and their ratings over the course of the show's runtime. I plotted episode ratings across the show's timeline and overlaid a linear line of best fit for each episode type. Using this chart, I aimed to explore how audience reception varied between MotW and plot-driven episodes over time, determine which episode type was generally better received, analyze how reception evolved throughout the series, and identify how each episode type performed by the end of the show's run.



```
ggplot(supernatural_episodes, aes(x = episode_id, y = imdb_episode_rating, color =
factor(mow_filler), group = mow_filler)) +
  geom_line(size = 0.5) +
  geom_point(size = 1) +
  geom_smooth(method = "lm", aes(color = factor(mow_filler)), se = FALSE, linetype = "dashed") +
  scale_color_manual(values = c("1" = "steelblue", "0" = "tomato3"), labels = c("Plot",
"MotW"))+
  theme_light() +
  labs(title = "Development of MotW VS Plot Episode Ratings over Runntime", x = "Episode
Number", y =  "Episode Rating")
```



This chart demonstrates that, overall, plot-driven episodes were consistently rated higher than MotW episodes throughout the show's runtime. However, the gap between the two began to narrow as the series approached its conclusion.

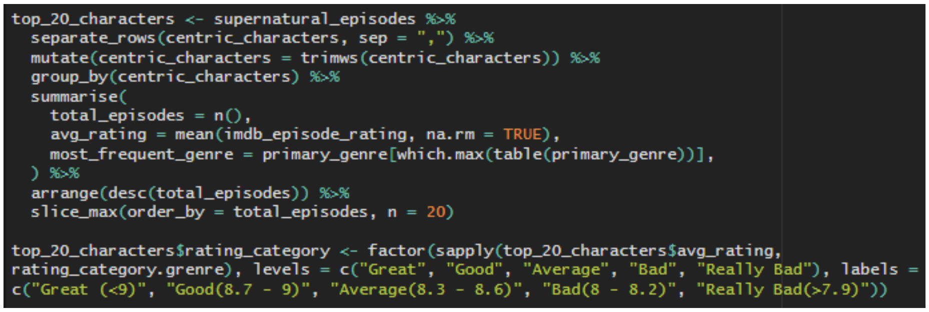
While both episode types show a downward trend in ratings over time, MotW episodes are declining at a slower rate compared to plot-driven episodes. With each season typically consisting of 40% to 60% MotW episodes, maintaining a similar balance in Season 16 seems like a solid strategy. This approach leverages the relative stability of MotW episodes while still incorporating the high-performing potential of plot-driven content.

It is possible that MotW episodes could surpass plot-driven episodes in ratings if the current trend continues into Season 16.

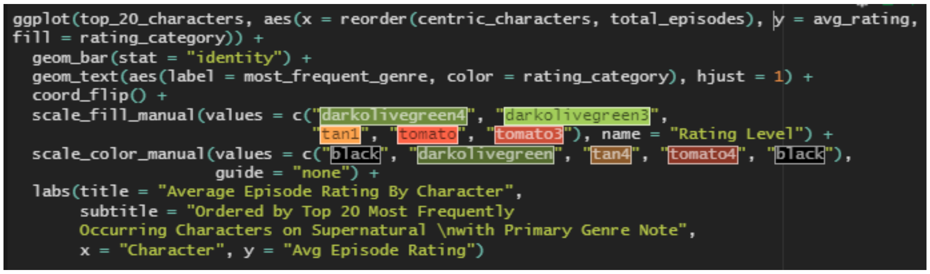
Visualization 6: Characters

With all my other questions answered, I turned to my final one: Which characters had the highest-rated episode average? The code for this analysis was the most complex of the group. It required splitting the character variable into separate rows, using the summarize() function to calculate the average episode rating and primary genre associated with each character, and then filtering for the top 20 characters based on the number of episode appearances.

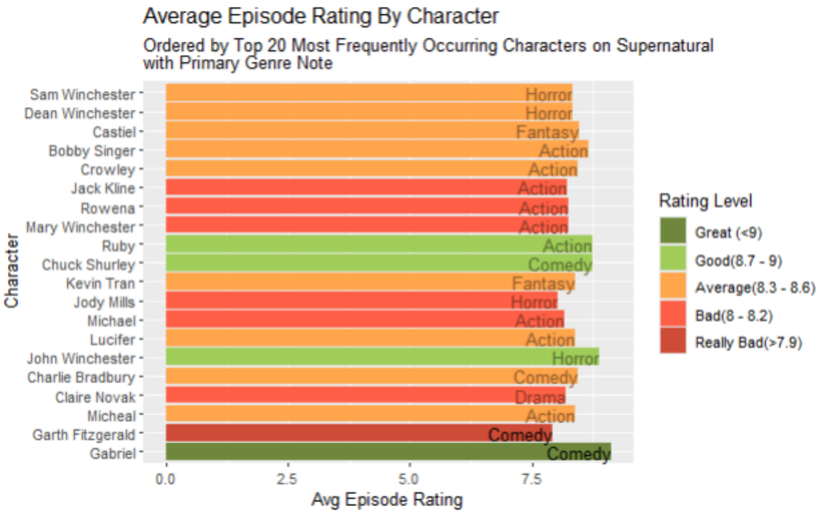
I visualized this data using a flipped coordinate bar chart, with the bars colored according to the episode rating categories.



```
top_20_characters <- supernatural_episodes %>%
  separate_rows(centric_characters, sep = ",") %>%
  mutate(centric_characters = trimws(centric_characters)) %>%
  group_by(centric_characters) %>%
  summarise(
    total_episodes = n(),
    avg_rating = mean(imdb_episode_rating, na.rm = TRUE),
    most_frequent_genre = primary_genre[which.max(table(primary_genre))],
  ) %>%
  arrange(desc(total_episodes)) %>%
  slice_max(order_by = total_episodes, n = 20)

top_20_characters$rating_category <- factor(sapply(top_20_characters$avg_rating,
rating_category.grenre), levels = c("Great", "Good", "Average", "Bad", "Really Bad"), labels =
c("Great (<9)", "Good(8.7 - 9)", "Average(8.3 - 8.6)", "Bad(8 - 8.2)", "Really Bad(>7.9)"))
```



```
ggplot(top_20_characters, aes(x = reorder(centric_characters, total_episodes), y = avg_rating,
fill = rating_category)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = most_frequent_genre, color = rating_category), hjust = 1) +
  coord_flip() +
  scale_fill_manual(values = c("darkolivegreen4", "darkolivegreen3",
                               "tan1", "tomato", "tomato3"), name = "Rating Level") +
  scale_color_manual(values = c("black", "darkolivegreen", "tan4", "tomato4", "black"),
                     guide = "none") +
  labs(title = "Average Episode Rating By Character",
       subtitle = "Ordered by Top 20 Most Frequently
       Occurring Characters on Supernatural \nwith Primary Genre Note",
       x = "Character", y = "Avg Episode Rating")
```



Interestingly, only one character achieved an overall average rating of more than 9.0: Gabriel. While he had the fewest episode appearances among the top 20 characters, this actually adds to his intrigue and potential, as he has many unresolved plot lines to offer a new season. Based on my previous findings, Gabriel emerges as a strong candidate for centering Season 16.

Gabriel is an interesting character in *Supernatural* because he embodies the archetype of a trickster. His episodes are known for their comedic elements that challenge the brothers' resolve in unconventional and entertaining ways. Furthermore, his story remains oddly unresolved. While he was seemingly killed on-screen in Season 5, later seasons revealed he was not as dead as we thought, simply trapped. In Season 9, there were hints that he may have escaped, but he was never seen again.

Given Gabriel's high episode ratings, his alignment with the comedic primary theme identified earlier, and the unresolved threads of his story, he feels like the perfect character to bring back as the central figure for Season 16.

## ## Conclusions

Season 15 of *Supernatural* left the series in a challenging narrative position: the two main characters, Sam and Dean Winchester, both died and coming to terms with their afterlives. Many other beloved characters (such as Chuck Shurley, Ruby, and John Winchester) were also dead, and the show's reputation for relying heavily on dramatic, last-minute resurrections to resolve conflicts has quickly grown old. To address these challenges while building on what has historically resonated with audiences, I propose the following outline for a hypothetical *Supernatural* Season 16.

Season 16 would follow Sam and Dean as they navigate their new existence in heaven, struggling to adjust to life after decades of turbulent adventures on Earth. Enter Gabriel, a mischievous and comedic figure whose chaotic antics often force the brothers to confront themselves and their unresolved pasts. Gabriel provides the perfect vehicle for tying up loose ends in their stories, presenting the brothers with a series of action-packed and humorous trials designed to test their resolve and help them and the audience find closure in the story.

The analysis of the *Supernaturanl's* previous seasons revealed that episodes with a strong comedic focus consistently outperformed other genres, making comedy an ideal overarching theme for this season. Gabriel, who had the highest-rated episodes of all recurring characters in the dataset, would take center stage, using his trickster nature to drive both humor and action throughout the story. The balance of MotW and Plot episodes was identified as not having much, if any, impact on season reception, so past strategies of having a roughly 50/50 split would be used to maintain steady pacing and variety. The decision to set the season in heaven also provides an opportunity to explore fresh storytelling while steering clear of overused resurrection tropes. Additionally, this setting aligns with audience preferences by utilizing fantasy elements.

With a foundation rooted in statistical insights, this proposed Season 16 strikes a balance between honoring *Supernat*ural's legacy and allowing the show to leave off on a better note than its current 6.9 series finale rating.

December 2nd, 2024 10:35pm supernatural