Team 9 Project Tech Bootcamp

Ethan Li, Sunday Nwanyim, Foster Mosden, Jenevive Zhang, Pacifique Iradukunda

1. What are the dimensions of the dataset (i.e. how many rows and columns)?
    a. There are 18 columns and 7,514,280 rows in the dataset.

```
133 •  SELECT COUNT(*) AS num_rows FROM reviews_clean;
134 •  SELECT COUNT(*) AS num_columns
135    FROM information_schema.columns
136    WHERE table_name = 'reviews_clean' AND table_schema = 'MSBA_Team9';
137
138
139
140
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

num_columns
18

```
132
133 •  SELECT COUNT(*) AS num_rows FROM reviews_clean;
134 •  SELECT COUNT(*) AS num_columns
135    FROM information_schema.columns
136    WHERE table_name = 'reviews_clean' AND table_schema = 'MSBA_Team9';
137
138
139
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

num_rows
7514280

2. How many unique products are there in the dataset?

There are 672,404 unique products in the dataset.

```
138 •  SELECT COUNT(DISTINCT product_id) as unique_products
139    FROM reviews_clean;
140
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

unique_products
672404

3. How many unique customers are there in the dataset?
   There are 3,819,430 unique customers in the dataset.

```
141 •  SELECT COUNT(DISTINCT customer_id) as unique_customers
142    FROM reviews_clean;
143
144
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

unique_customers
3819430

4. How many unique combinations of product_id and product_parent are there?

There are 672,859 unique combinations of product_id and product_parent.

```
144 •  SELECT COUNT(DISTINCT product_id, product_parent) AS unique_combinations
145    FROM reviews_clean;
146
147
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

unique_combinations
672859

5. *What is the average product rating? If you partition the reviews by whether the person's purchase was verified, does the average product rating differ between the two groups? Use a visualization to support your findings.*

```
use MSBA_Team9;
# create a new smaller table for question 5. we only need star_rating and verified_purchase
to answer
CREATE TABLE question_5 AS
SELECT star_rating, verified_purchase
FROM reviews_clean;
# Get the average of all star ratings in the table
SELECT AVG(star_rating) FROM question_5;
# Get the avg star ratings for each verified_purchase value
SELECT verified_purchase, AVG(star_rating) as avg_star_rating
FROM question_5
GROUP BY verified_purchase;
```

AVG(star_rating)
4.1849

verified_purchase	avg_star_rating
1	4.2397
0	3.8654

```
library(dplyr)
library(tidyverse)
library(magrittr)

# load in our filepath and then save the CSV as a dataframe
data_file <- "our filepath"
df_reviews <- read_csv(data_file)

# Map verified_purchase values to clear labels
df_reviews$verified_purchase <- factor(df_reviews$verified_purchase, levels = c(0, 1), labels
= c("Not Verified", "Verified"))

# create a ggplot using the df_reviews
# verified purchase status as the x axis, mean star rating as the y axis
ggplot(df_reviews, aes(x = verified_purchase, y = star_rating, fill = verified_purchase)) + #
set x and y axis, and align bar color with verified status
  geom_bar(stat = "summary", fun = "mean") + # use mean as summary statistic
  labs(x = "Verified Purchase", y = "Average Star Rating") + # add labels
  theme_minimal() + # use minimal theme to make things less visually busy
  coord_cartesian(ylim = c(0,5)) # set the legend of the y axis to 0-5
```

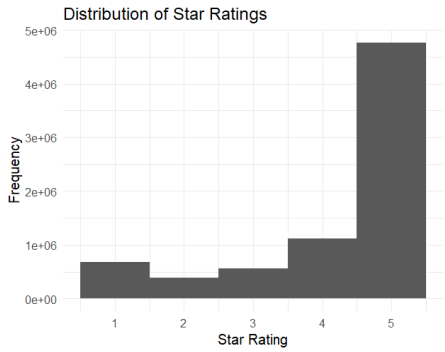
We found the average star rating across all reviews in any category to be 4.1849.
When you partition the data by whether or not the reviews are verified, you see that the average product rating is much higher for verified purchases (4.24) when compared to non-verified purchases (3.87).
Perhaps verified purchases are more likely to be people who bought the product for themselves, therefore having a better understanding of the product before using it, which might result in greater satisfaction.

6. What does the distribution of ratings look like? Use one or more visualizations to support your findings.

```
use MSBA_Team9;
# create a new smaller table for question 6. we only need star_rating
CREATE TABLE question_6 AS
SELECT star_rating
FROM reviews_clean;
# Get the count total number of reviews for each of the distinct possible
star_rating values
SELECT star_rating, COUNT(star_rating) AS
num_of_reviews
FROM question_6
GROUP BY star_rating
ORDER BY star_rating DESC;
```

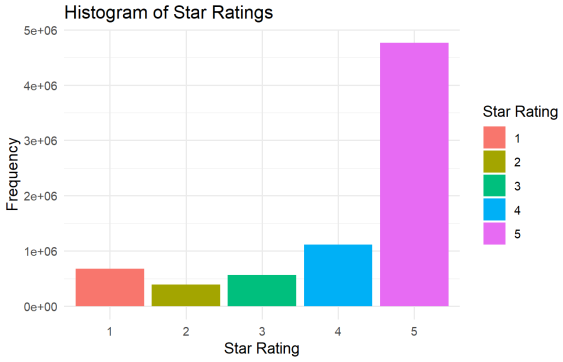
star_rating	num_of_reviews
5	4767589
4	1112452
3	565866
2	393006
1	675367

```
library(dplyr)
library(tidyverse)
library(magrittr)

# load in our filepath and then save the CSV as a dataframe
data_file <- "our filepath"
df_reviews <- read_csv(data_file)

# create ggplot histogram to visualize distribution of star_ratings
ggplot(df_reviews, aes(x = star_rating)) +  # specify dataframe and x axis
geom_histogram(binwidth = 1) +  # histogram with each bar covering 1 rating
  labs(x = "Star Rating", y = "Frequency", title = "Distribution of Star
Ratings") +  # labels and title
  theme_minimal()  # visually clean
  coord_cartesian(ylim = c(0,5000000)) # set y limit to 5,000,000
```

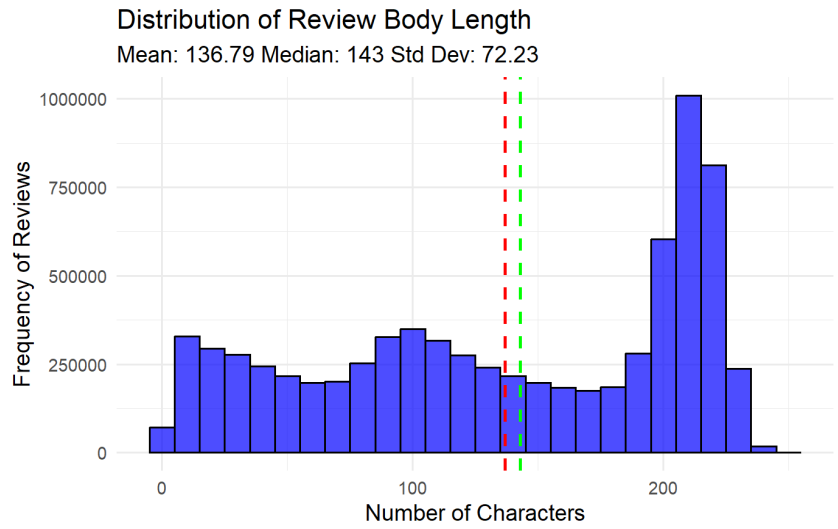


The distribution of ratings is heavily skewed left. It seems that there is a bias of customers to underutilize the 5-star rating scale, as the most common ratings are 5, 4, then 1. Perhaps customers who have any dissatisfaction with the product have a tendency to give it the lowest possible score in order to express their feelings, while customers who are in any way satisfied with the product are likely to give it a 4-5 because they have few issues with it.

7. For each customer, devise a metric that captures the “fairness” or “bias” of a reviewer. It is completely up to you to decide how to define this metric, and it can be either categorical or numerical. In your report, explain how you came up with your metric and why you think it makes sense to define it in this way. Then, using your metric, determine if the distribution of rating changes after you exclude the “unfair”/“biased” reviewers. Use one or more visualizations to support your findings.



My initial hypothesis was that shorter body reviews, as in a smaller number of character lengths, are more biased. It could be possible that in the event of just wanting to uplift and downgrade average reviews, someone put a high or low star rating and did not care to explain why. We can see that this conclusion is somewhat supported by the fact that the number of star ratings is really high for the 5-star rating. In contrast, even though the average character length for a star rating of 5 is significantly lower than the rest of the other ratings. The same can be said for a star rating of 4 and 1. To further explore this hypothesis we will analyze the distribution of the number of characters in each review based on star rating. Then based on the distribution of the number of characters we will identify the variance around the mean to see if the distribution is normal and see which number of characters falls outside 2SD of the mean.



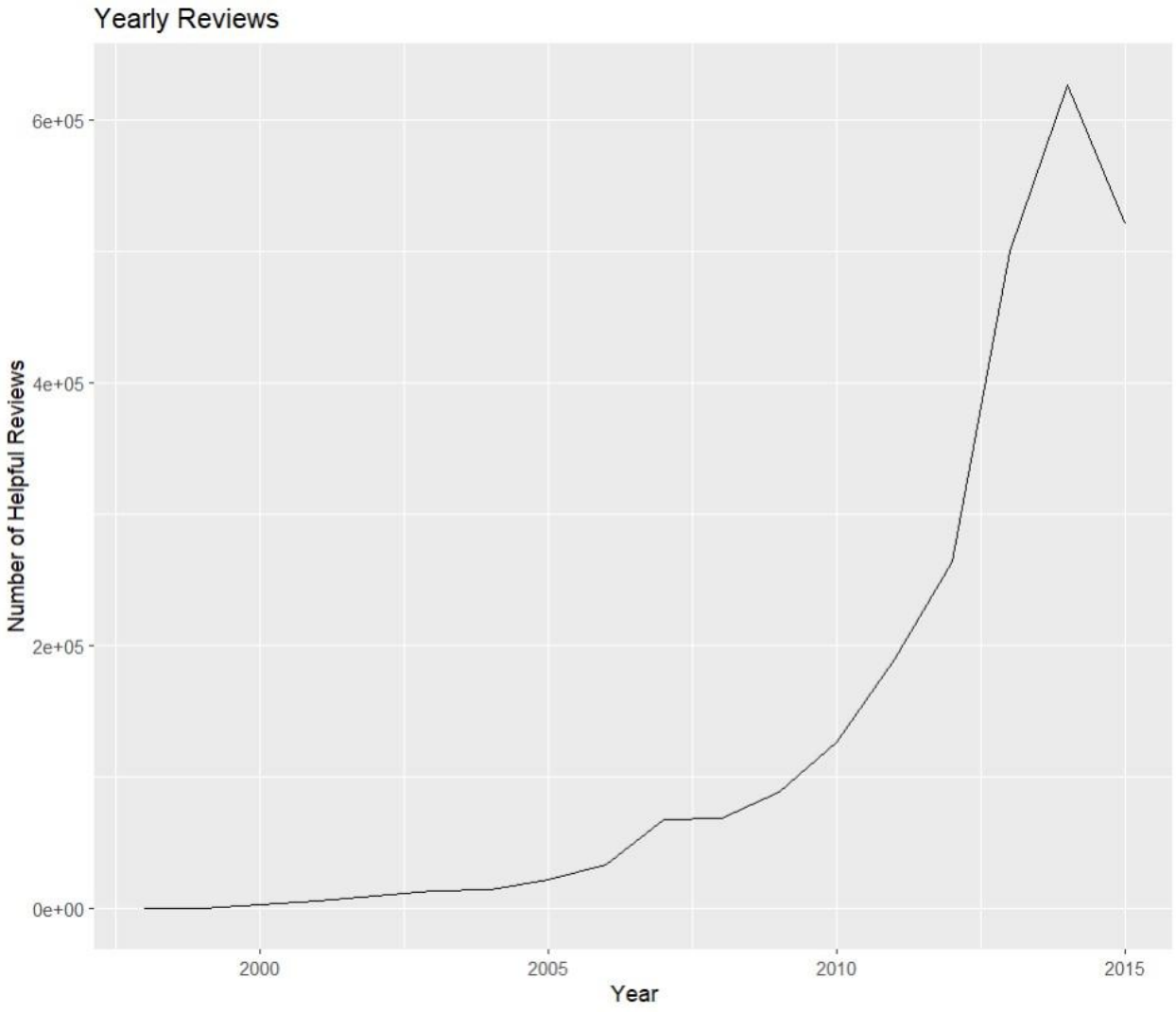
The distribution of the number of characters vs the frequency of reviews is shown above. Since the distribution is not normal, I have decided to use the Median absolute deviation (MAD) which is a robust measure of the variability in univariate data. Any values that fall outside our 2.5 MAD

from the median will be considered unusual. But we focus mainly on those that fall 2.5 outside of the median as we are focusing on short body reviews with no justification.

8. Aggregate the number of reviews to the monthly level. Then, determine whether the number of reviews exhibits seasonality within a calendar year. Describe what you find and use one or more visualizations to support your findings.

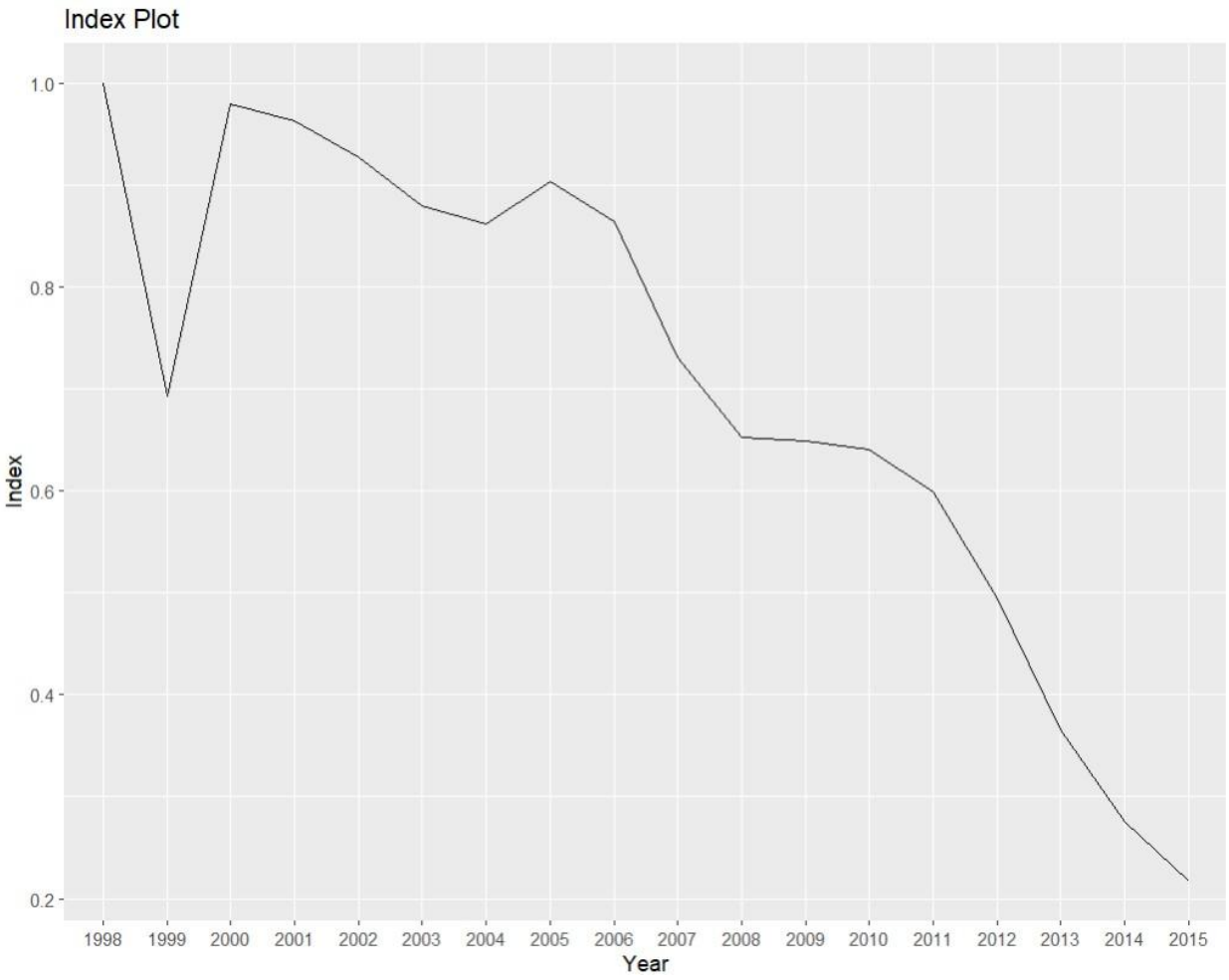
We can find evidence of seasonality in most of December and January. This can be attributed to Christmas and New Year holidays. The other months don't have a specific pattern.

9. Plot a time series with the year on the x-axis and the number of helpful reviews on the y-axis. Describe the plot and include the visualization in your report. A helpful review is defined as a review that has at least one helpful vote.



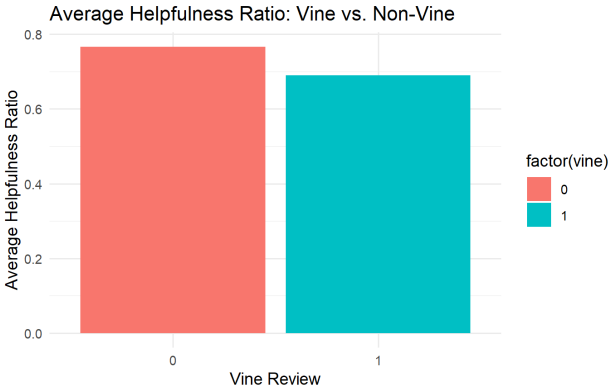
From the graph above we can see that the number of helpful reviews increased between year 2000 till 2015. From 2000 to 2008 there was a slow increase in the number of helpful reviews. From 2010 the number of helpful reviews exponentially increased until the year 2014. And the reviews experienced a little decline in 2015. Overall there has been an increase in helpful reviews.

10. Plot a time series with the year on the x-axis and an index representing the “adjusted” number of helpful reviews on the y-axis. The index at time t equals, where is the total number of $Nt\ Tt\ Nt$ helpful reviews for year t and is the total number of reviews for year t. Note that the totalTt the number of helpful reviews is not the same as the total number of helpful votes. Describe the plot and how it changes how you interpret the plot from Question 9. Include all visualizations in your report.

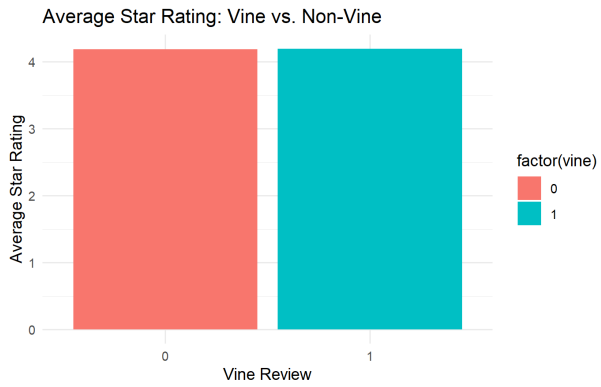


The plot in this question is the inverse of the plot in question 9. The plots start from a high in 1998 and then keep declining until 2015.

11. Based on this data, if Amazon.com asked you to help them decide whether to retain or discontinue the Vine program, what would you recommend to them? Explain and justify your reasoning. Use one or more visualizations to support your findings.



In this first visualization, we calculated an average helpfulness ratio by dividing the number of Vine reviews by the total number of reviews. Since vine reviews have a higher helpfulness ratio on average, it suggests that Vine reviewers provide higher-quality reviews that are more beneficial to other customers of helpful votes over the total number of votes separated by vine vs non-vine review.



Also, based on this visualization, which maps the average star rating of vine vs non-vine members. The average ratings are generally the same which indicates that there is no bias in the program since on average the reviews coming from members and those who aren't are roughly the same. Since on average vine reviews give more helpful reviews and due to the fact that on average vine and non-vine program reviews have the same rating, the program should be continued.

12.

Identifying the most feasible approach in data cleaning processing was challenging due to the diversity & volume of data. Our solution was to evaluate the inconsistencies that occurred in the database and customized certain functions in order to address the issues. Additionally, we've discovered some commands weren't successfully operating on some group mates' SQL workbench. We had to simultaneously debug the workarounds, such as meeting in person & zoom: in order to complete the assignment compatibly. The safe update mode has restricted us from performing & modifying certain data, which lead us in setting the safe update to zero in order to be authorized to continue with the cleaning process. In conclusion, our team has achieved efficiency in the project by distributing duties based on each group mate's strength.

13.

The supportive nature of our group has made our learning curve rewarding, we've realized the value of teamwork and collaboration when it comes to complicated concepts of data or any knowledge. Throughout the boot camp, we've had a profound message that the power of data lies in the application. The coursework has built a profound foundation for us in utilizing data as a tool for the corporate world.