Portfolio Sample:
New York Times Crossword Data Analysis

Olivia Grunseich
October, 2023

Spiral Escalator LLC

Introduction

Welcome to this unique exploration where my passion for **puzzles** and **data** converge. This case study serves as a showcase of how seemingly disparate interests can harmoniously intertwine. In my journey, I've discovered that the intricate patterns of crossword puzzles and the meticulous analysis of data share a common thread—a fascination with deciphering complex information. Here, we'll delve into a fusion of these worlds, demonstrating how the art of crosswords and the science of data patterns come together to reveal fascinating insights.

New York Times Crossword Analysis

Can analyzing the New York Times Crossword data from the past 30 years uncover strategies and patterns that are useful for completing new puzzles?

Background

New York Times Crossword

- **Grid Structure:** The New York Times crossword consists of a square grid, typically 15x15 squares for daily puzzles, with black squares dividing it into sections.
- **Clues:** Each puzzle has a set of clues, both Across and Down, corresponding to grid squares. Clues are typically short, cryptic phrases or definitions that hint at the answer.
- **Word Intersections:** Solving the puzzle entails placing words in the grid so that they share common letters and intersect accurately with other words, creating an interconnected network of solutions.
- **Difficulty Levels:** Puzzles vary in difficulty, from easy to challenging. The easiest puzzle is Monday, increasing in difficulty through Saturday. Sunday puzzles are larger than the other days of the week, typically with a difficulty similar to that of a Thursday puzzle.

Background

Data

- The original data set used for this study can be found **here.**
- The data consists of all clues and answers for The New York Times crossword from **11/21/93** through **10/31/21**
- Each entry in the data set consists of the crossword **date**, **word**, and **clue**
- There are **781,573** entries in total

Field	Description
Date	The publishing date of the crossword puzzle
Word	The word(s) or phase(s) entered into the puzzle grid, capitalized without spaces. “The answer”
Clue	The published hint that prompts the solver’s answer. “The clue”

Date	Word	Clue
10/31/2021	PAT	Action done while saying "Good dog"
10/31/2021	RASCALS	Mischief-makers
10/31/2021	PEN	It might click for a writer
10/31/2021	SEP	Fall mo.
10/31/2021	ECO	Kind to Mother Nature

Sample entries of data set

Goals

What questions are we looking to answer using the data available?

- What are the most common crossword answers?
- What are the most common crossword clues?
- In what order should a solver approach crossword clues?

Data Cleaning and Setup using R

Steps

The computer programing language R was used for all analysis and graphic visuals in this presentation. For the specific code used, refer to the appendix

1. **Updated *date* column name to *nytc_date* and converted dates in character format to date format**
   The word "date" is a common word used in R functions. To avoid confusion, the column name was changed to *nytc_dates*. The entries in this column initially registered as characters rather than actual date values. The format was updated to fix this.

2. **Added new column for day of the week (*day_of_week*)**
   Using the information from the formatted *date* column, we can establish the day of the week for each entry

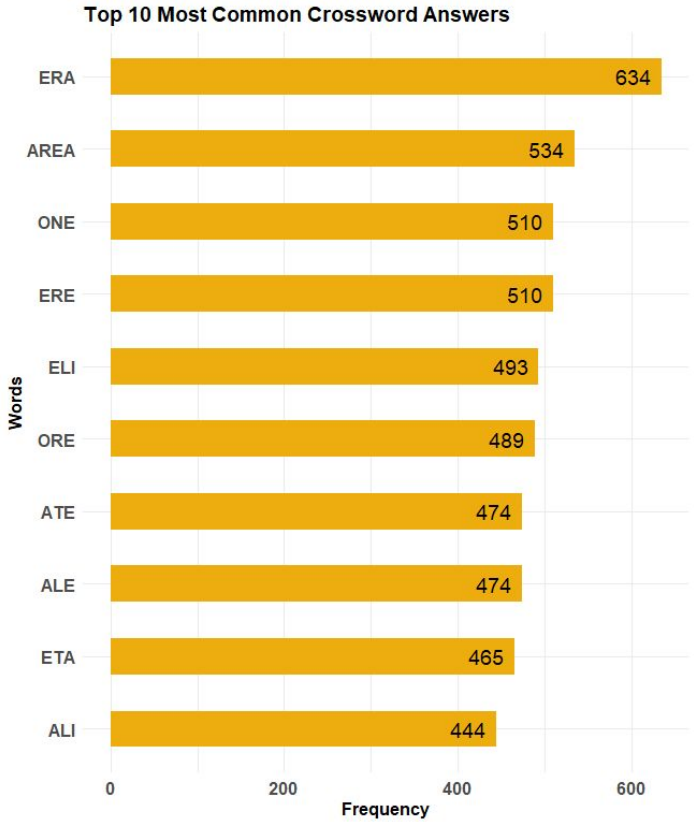
3. **Added column for word length (*word_len*)**
   This new column shows the number of letters in each *word* entry

4. **Searched for and deleted all entries that were less than 3 letters in length, violating NYT crossword rules**
   4 entries were deleted, changing the total number of entries to 781,569

nytc_date	Word	Clue	word_len	day_of_week
2021-10-31	PAT	Action done while saying "Good dog"	3	Sunday
2021-10-31	RASCALS	Mischief-makers	7	Sunday
2021-10-31	PEN	It might click for a writer	3	Sunday
2021-10-31	SEP	Fall mo.	3	Sunday
2021-10-31	ECO	Kind to Mother Nature	3	Sunday

Sample entries of cleaned data set

What are the most common crossword answers?

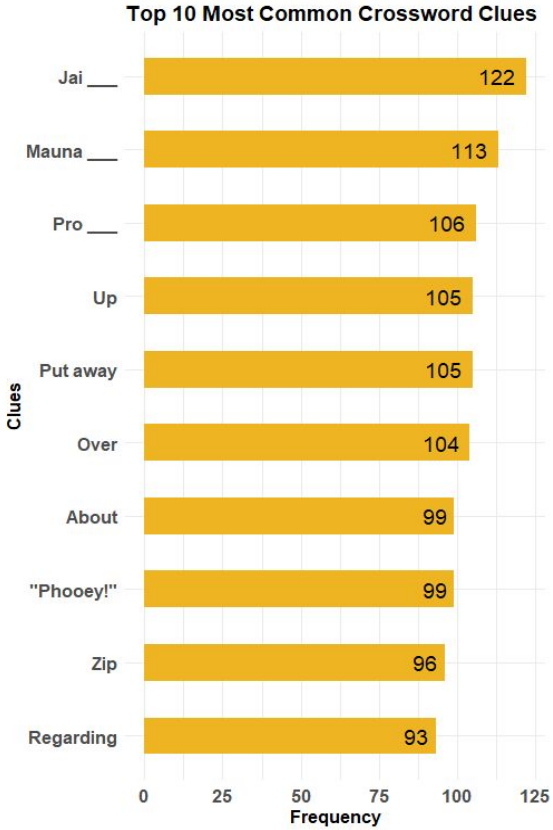


- When you look at the top words used from 1993 to 2021, **"ERA"** is the clear winner, having been used 634 times, a whole 100 above the second place answer, **"AREA"**

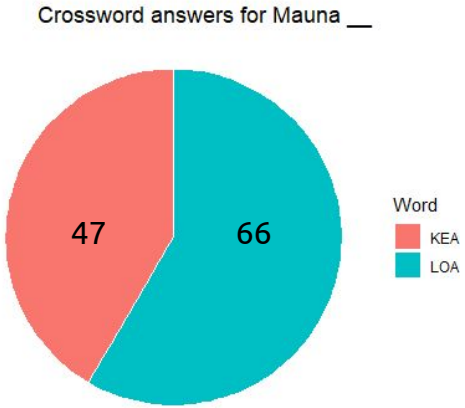
- It's clear from this list that the most common words used are generally short (9 out of the top 10 are **3-letter words**) and **vowel-heavy** (all of the top 10 have at least 2 vowels)

- This makes sense if we think about the structure of these puzzles; short words like these are **needed to construct the longer words and phrases**

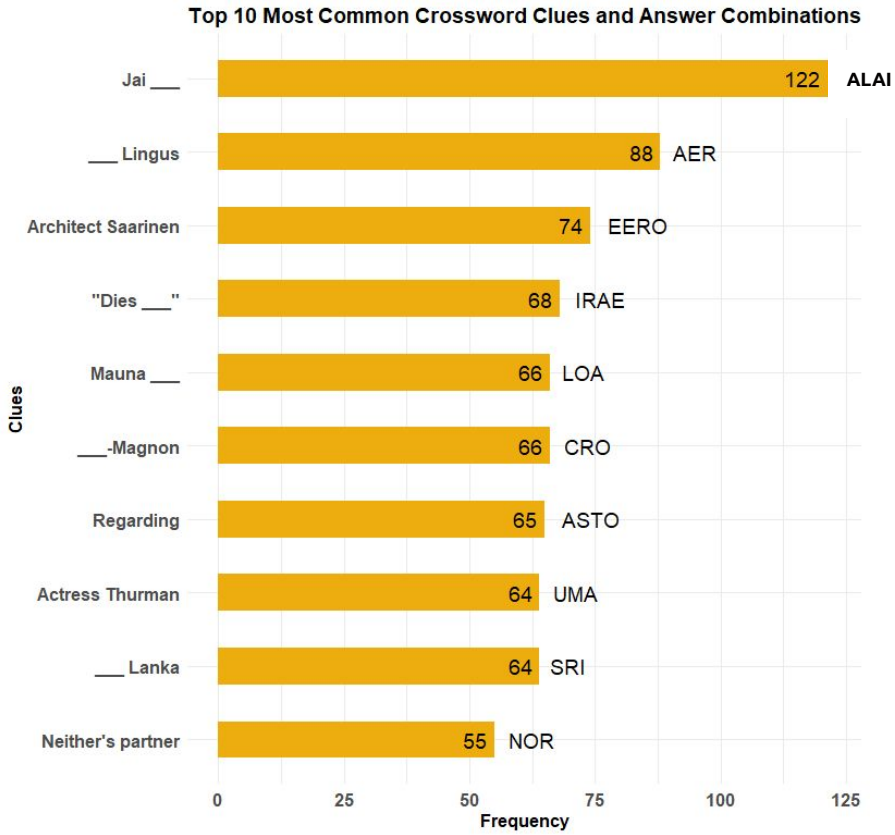
What are the most common crossword clues?



- When we look at most-used clues, one might assume that they might match the top words, but this is not the case. In fact, **none of the top 10 words correspond to the top 10 clues**
- This is because crossword builders do not like to reuse clues. For a popular word like "ERA" there could be as many as 30 different clues associated with it
- Additionally, clues are often used to misdirect puzzlers, leading them to potentially wrong answers. For example, the popular clue **Mauna___** is associated with 2 different answers: **"KEA"** or **"LOA"**. The correct choice is dependent on the puzzle



What are the most common crossword clues? Continued



- When we look at clue/answer combinations, we see some of the top individual clues, but not all

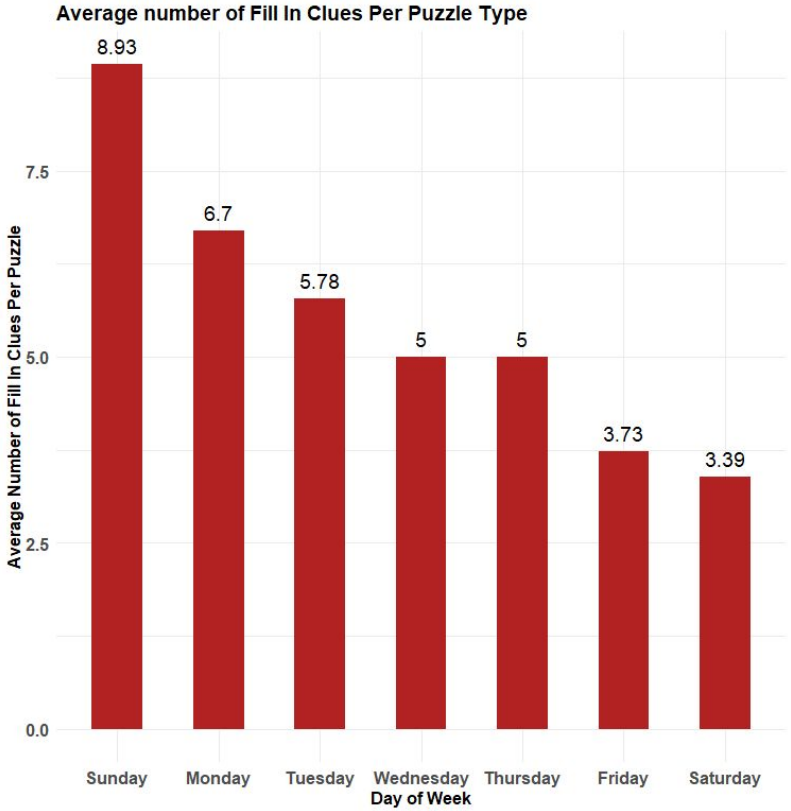
- Similar to the most-used words, these answers are mostly all **short** and **vowel-heavy**. However, these words are **more specific** to the clues that are provided with them, meaning you will not likely see other clues associated with these words

- 6 out of the top 10 clue/answer combinations have a fill in the blank clue. This begs the question: **Are fill in the blank clues easier to solve than other clues?**

In what order should a solver approach crossword clues?

Are fill in the blank clues easier to solve than other clues?

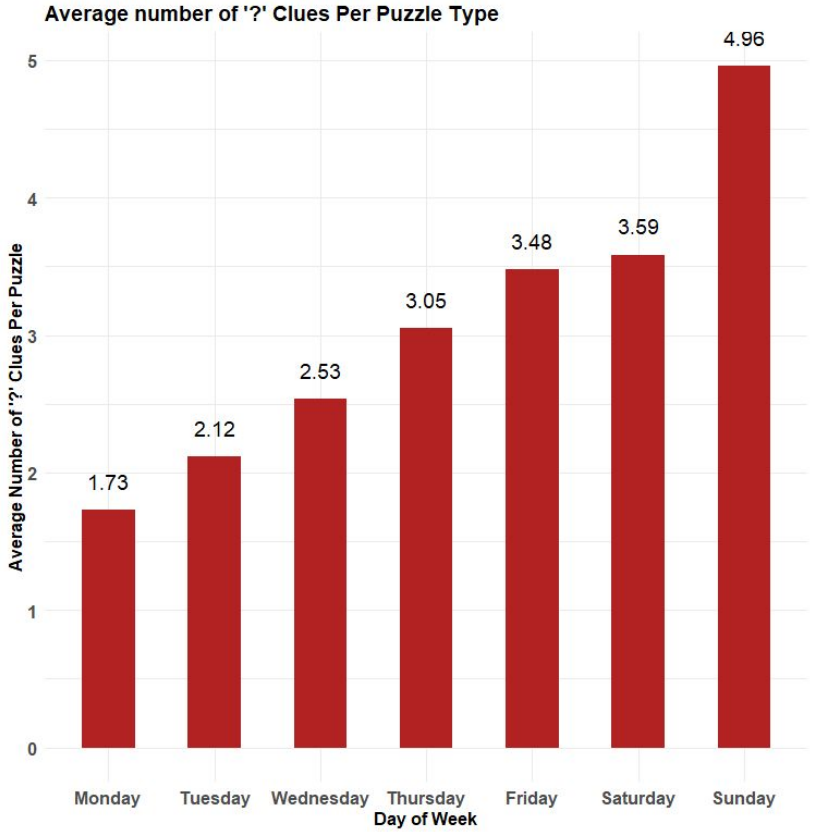
- Knowing that the **puzzle difficulty increases from Monday through Saturday**, all fill in the blank clues were filtered and the average number of this clue type used per puzzle was determined
- As assumed from the previous slide, **fill in the blank clues are used more often in the easier puzzles**, decreasing in use for the more difficult days of the week
- Sundays are considered an outlier in this study because Sunday puzzles are larger in size and would therefore have a larger number of clues overall, skewing the average



In what order should a solver approach crossword clues? Cont

Are “?” clues harder to solve than other clues?

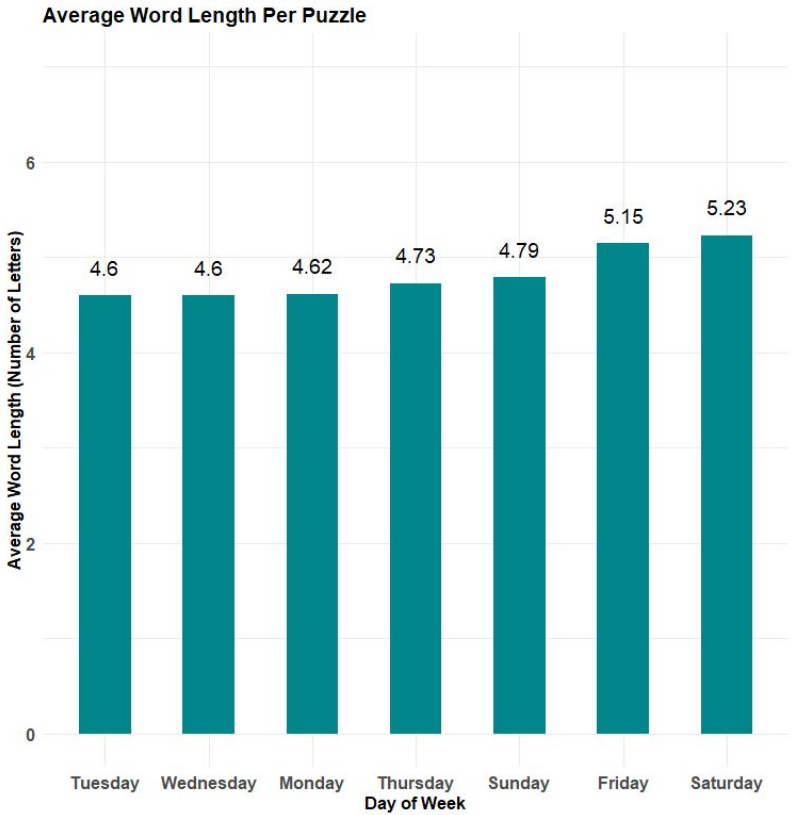
- A clue with a **"?"** often indicates a playful or punny hint, where the answer is not meant to be taken literally, adding an element of **wordplay** to the puzzle
- An analysis was done to see if these clues also had a correlation with puzzle difficulty
- As suspected, **“?” clues were less common in the easier puzzles** and increased as the week progressed, likely because these clues require the user to **think outside of the box**
- Again, Sunday should be dismissed in this analysis because of its grid size

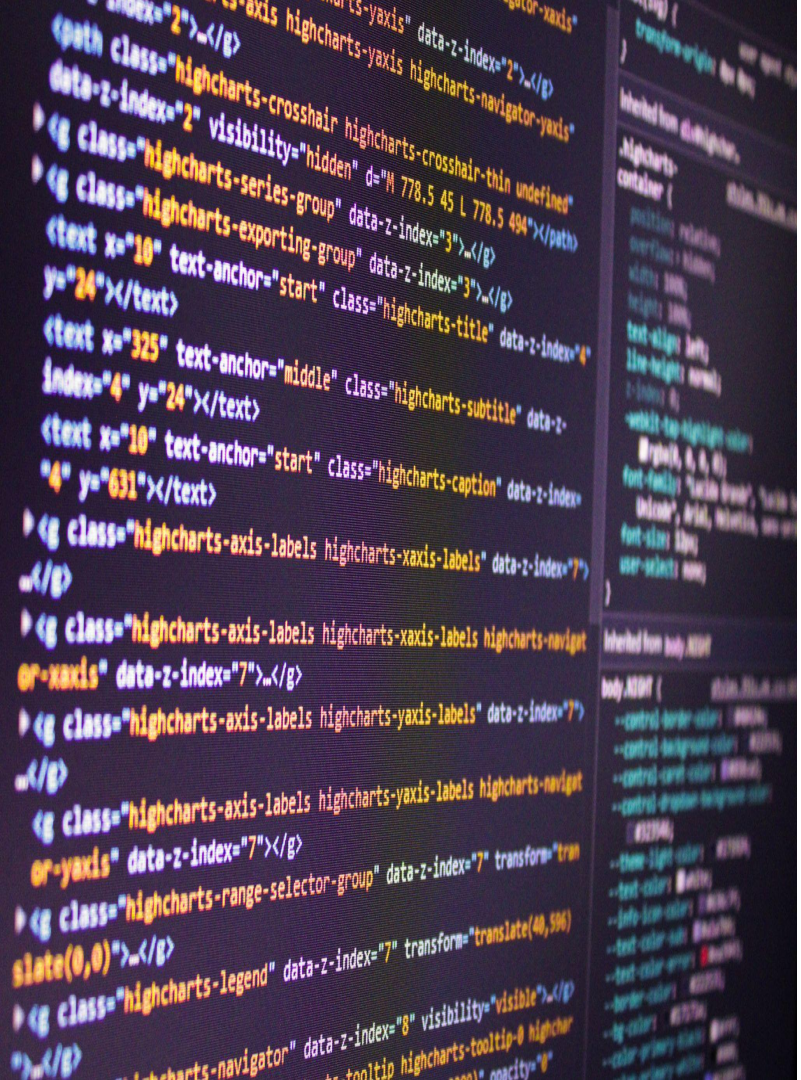


In what order should a solver approach crossword clues? Cont

Are smaller words easier to solve than larger words?

- From the previous analysis on the most common crossword answers, we know that **3-letter words have the highest frequency of being reused**
- When we look at **average word length** based on puzzle difficulty, it's not as predictable as the previous studies
- Unexpectedly, **Monday, Tuesday,** and **Wednesday** all have a **lower average** of about 4.6 letters, but the averages are all so close that **Monday** is actually the **highest of the 3**
- **Thursday, Friday** and **Saturday** are in the order you would suspect, **increasing in length throughout the week**
- However, **Sunday** surprisingly falls in between the later days of the week, taking the **5th** spot despite being a larger puzzle overall. This could support the argument that **Sunday's difficulty is similar to that of a Thursday puzzle** because it likely has many lower letter answers that bring down its average





In what order should a solver approach crossword clues?

Data-driven strategy

1. **Short (3-5 letters) words should be attempted first.** These should be the easiest to solve and will help fill in the larger words

2. **Fill in the blank clues should be attempted earlier**. These should be easier to solve than other clue types

3. **Don't spend too much time on "?" clues in the beginning**. These are more difficult to solve

4. **Save long clues for later.** These tend to be more difficult and will be easier if you have crossed clues filled in already

Appendix

R Code Used for Analysis

```
##DOWNLOAD AND INSTALL
#Install all data and packages needed to analyze

nytcrosswords <- read.csv("C:/Program Files/R/Olivia/nytcrosswords.csv")
 View(nytcrosswords)

install.packages("ggpmisc")
install.packages("dplyr")
install.packages("tidyverse")

library(ggpmisc)
library(tibble)
library(dplyr)
library(quantreg)
library(lubridate)
library(stringr)
update.packages()


##CLEAN DATA
#Check data types of file

head(nytcrosswords)
#Date column is type "character"

#Changes the name of the column to nytc_dates
colnames(nytcrosswords)[1] <- "nytc_date"
#could have also used rename(nytcrosswords, nytc_date = date)

#The below converts each column into an individual dataframe for individual analysis (this is not needed but good to know)
nytc_date <- nytcrosswords[1]
nytc_word <- nytcrosswords[2]
nytc_clue <- nytcrosswords[3]

#Convert dates in character format to date format
nytc_date <- as.Date(nytcrosswords$Date,"%m/%d/%Y")

#Add new column for days of the week
nytcrosswords$day_of_week = weekdays(nytcrosswords$nytc_date)
# The below adds day of week as a dataframe
# day_of_week <- weekdays(nytcrosswords$nytc_date)

#Add a column for word length
nytcrosswords$word_len = str_length(nytcrosswords$Word)


#check to see if there are words that break the rule that there should be at least 3 letters per word
nytcrosswords %>%
  +     filter(word_len == 2)
nytcrosswords %>%
  +     filter(word_len == 1)
nytcrosswords %>%
  +     filter(word_len == 0)
#This shows us that there are (3) 2-letter words, (1) 1-letter word, and (0) 0-letter words

#To clean this, we make a new dataframe without those values
nytc_clean <- subset(nytcrosswords, word_len > 2)
#We know that it works because the nytc_clean set has 4 less values than the original set
```

```
##INITIAL DATA SUMMARY

count(nytcrosswords, Clue)
#This shows us that there are about 462k different clue/word pairings in this data set, meaning the rest are duplicates


##MOST COMMON ANSWERS
#Count how often each word is used
common_ans <- nytc_clean %>%
  + count(`Word`, sort = TRUE)

#Change "n" column to "Frequency"
colnames(common_ans)[2] <- "Frequency"

#Filter for the top 10 answers
common_ans_top <- top_n(common_ans, n=10)

#Table showing 10 most common words
common_ans_top %>%
  ggplot(aes(x=Frequency, y=reorder(Word, Frequency)))+
  geom_bar(stat = 'identity', width = 0.5, fill = "darkgoldenrod2")+
  geom_text(aes(label = Frequency), hjust=1.3, color = "black",size = 5)+
  theme_minimal()+
  labs(title = "Top 10 Most Common Crossword Answers", face="bold", y= "Words")+
  theme(plot.title = element_text(size = 14, face= "bold"),
        axis.title.x = element_text(size = 12, face = "bold"),
        axis.title.y = element_text(size = 12, face = "bold"),
        axis.text.x = element_text(size = 12, face = "bold"),
        axis.text.y = element_text(size = 12, face = "bold"),)



#MOST COMMON CROSSWORD CLUE WITHOUT ANSWER
#Count how often each word is used
common_clue <- nytc_clean %>%
   count(`Clue`, sort = TRUE)

#Change "n" column to "Frequency"
colnames(common_clue)[2] <- "Frequency"

#Filter for the top 10 answers
common_clue_top <- top_n(common_clue, n=10)

#Graph showing 10 most common clues
common_clue_top %>%
  ggplot(aes(x=Frequency, y=reorder(Clue, Frequency)))+
  geom_bar(stat = 'identity', width = 0.5, fill="goldenrod2")+
  geom_text(aes(label = Frequency), hjust=1.3, color = "black", size=5)+
  theme_minimal()+
  labs(title = "Top 10 Most Common Crossword Clues", face="bold", y= "Clues")+
  theme(plot.title = element_text(size = 14, face= "bold"),
      axis.title.x = element_text(size = 12, face = "bold"),
      axis.title.y = element_text(size = 12, face = "bold"),
      axis.text.x = element_text(size = 12, face = "bold"),
      axis.text.y = element_text(size = 12, face = "bold"),)


#Mauna___ example
Mauna <- filter(nytc_clean, Clue =="Mauna ___")

#Mauna Bar graph (not used)
Mauna %>%
  ggplot(aes(x=Word))+
  geom_bar(width = .5)+
  geom_text(aes(label = after_stat(count)),stat = "count",vjust=1.5, color = "white")+
  theme_minimal()+
  labs(title = "Crossword answers for Mauna __", y= "Frequency")

#Mauna Pie Chart
Mauna2 <- Mauna %>%
  count(`Word`, sort = TRUE)

#Making the Pie Chart
Mauna2 %>%
  ggplot(aes(x="", y=n, fill=Word))+
  geom_bar(stat = "identity",width = 1, color="white")+
  coord_polar("y")+
  theme_void()+
  labs(title = "          Crossword answers for Mauna __", y= "Frequency")



##MOST COMMON CROSSWORD CLUE/ANSWER COMBO
#Count how often each clue/answer combo is used
common_clue_combo <- nytc_clean %>%
  + count(`Word`,`Clue`, sort = TRUE)

#Change "n" column to "Frequency"
colnames(common_clue_combo)[3] <- "Frequency"

#Filter for the top 10 answers
common_clue_combo_top <- top_n(common_clue_combo, n=10)

#Graph showing 10 most common words
common_clue_combo_top %>%
  ggplot(aes(x=Frequency, y=reorder(Clue, Frequency)))+
  geom_bar(stat = 'identity', width = 0.5, fill= "darkgoldenrod2")+
  geom_text(aes(label = Frequency), hjust=1.3, color = "black", size= 5)+
  geom_text(aes(label = Word), hjust=-0.3, color = "black", size= 5)+
  theme_minimal()+
  labs(title = "Top 10 Most Common Crossword Clues and Answer
Combinations",face="bold", y= "Clues")+
  theme(plot.title = element_text(size = 14, face= "bold"),
        axis.title.x = element_text(size = 12, face = "bold"),
        axis.title.y = element_text(size = 12, face = "bold"),
        axis.text.x = element_text(size = 12, face = "bold"),
        axis.text.y = element_text(size = 12, face = "bold"),)
```

```
## FILL IN THE BLANK CLUES
#Filter for "__"
fill_in_clues <- nytc_clean %>%
  filter(grepl('__',Clue,fixed = TRUE))

#Count how many "___" clues there are
fill_in_clues_table <- fill_in_clues %>%
  group_by(day_of_week)%>%
  count(nytc_date,day_of_week) %>%
  summarise(mean(n))

#Graph showing avg fill in clues per puzzle
fill_in_clues_table %>%
  ggplot(aes(x= reorder(day_of_week,-`mean(n)`), y=`mean(n)`)) +
  geom_bar(stat = 'identity', width = 0.5, fill="firebrick") +
  geom_text(aes(label = round(`mean(n)`,2)), vjust=-0.7, color = "black", size=5)+
  theme_minimal() +
  labs(title = "Average number of Fill In Clues Per Puzzle Type",y= "Average Number of Fill In Clues Per Puzzle", x="Day of
Week")+
  theme(plot.title = element_text(size = 14, face= "bold"),
        axis.title.x = element_text(size = 12, face = "bold"),
        axis.title.y = element_text(size = 12, face = "bold"),
        axis.text.x = element_text(size = 12, face = "bold"),
        axis.text.y = element_text(size = 12, face = "bold"),)
## in future, try to figure out avg of fill in clues compared to all clues from that day to make sunday on the same plane as
the others


##? CLUES
#Filter for "?"
question_clues <- nytc_clean %>%
  filter(grepl('?',Clue,fixed = TRUE))

#Count how many "?" clues there are
question_clues_table <- question_clues %>%
  group_by(day_of_week)%>%
  count(nytc_date,day_of_week) %>%
  summarise(mean(n))

#Graph showing avg "?" clues per puzzle
question_clues_table %>%
  ggplot(aes(x= reorder(day_of_week,`mean(n)`), y=`mean(n)`)) +
  geom_bar(stat = 'identity', width = 0.5, fill="firebrick") +
  geom_text(aes(label = round(`mean(n)`,2)), vjust=-1.3, color = "black", size=5)+
  theme_minimal() +
  labs(title = "Average number of '?' Clues Per Puzzle Type",y= "Average Number of '?' Clues Per Puzzle", x="Day of
Week")+
  theme(plot.title = element_text(size = 14, face= "bold"),
        axis.title.x = element_text(size = 12, face = "bold"),
        axis.title.y = element_text(size = 12, face = "bold"),
        axis.text.x = element_text(size = 12, face = "bold"),
        axis.text.y = element_text(size = 12, face = "bold"),)
```

```
##AVERAGE WORD LENGTH PER DAY OF WEEK
#Make table summarizing the average word length
word_len_table <- nytc_clean %>%
  group_by(`nytc_date`,`day_of_week`) %>%
  summarise(avg_len = mean(word_len)) %>%
  group_by(day_of_week) %>%
  summarise(avg_len2 = mean(avg_len))

#Another analysis but we will use the above
nytc_clean %>%
  +     group_by(day_of_week) %>%
  +     summarise(avg_word_len = mean(word_len))

#Graph for average word length
word_len_table %>%
  ggplot(aes(x= reorder(day_of_week,avg_len2), y=avg_len2)) +
  geom_bar(stat = 'identity', width = 0.5, fill="turquoise4") +
  geom_text(aes(label = round(`avg_len2`,2)), vjust=-1.3, color = "black", size=5)+
  theme_minimal() +
  ylim(0,7) +
  labs(title = "Average Word Length Per Puzzle",y= "Average Word Length (Number of
Letters)", x="Day of Week")+
  theme(plot.title = element_text(size = 14, face= "bold"),
        axis.title.x = element_text(size = 12, face = "bold"),
        axis.title.y = element_text(size = 12, face = "bold"),
        axis.text.x = element_text(size = 12, face = "bold"),
        axis.text.y = element_text(size = 12, face = "bold"),)
```