ggplot2 Practice

MAS 627

Market Share in Iowa

You work for Proximo Spirits, a tiny liquor distributor that is battling for a piece of the pie in Iowa. You were brought in to oversee the account managers in November of last year. Historically, Proximo has held about 2.5% market share in Iowa, and you set the ambitious goal of crossing over 3% this year.

The year is halfway over and it's time to check progress toward our goal!

NOTE: Data runs from July 2021 through June 2022, so "November of last year" means November 2021.

Instructions

Starter plot

1. Fix the date variable, and year and month to the data.
2. Filter for our company and calculate total sales by month
3. Visualize this with a barplot

Convert to percentages (NEED: Our Sales / Total Sales)

4. Create a new dataset that stores monthly totals
5. Join this in with your current dataset
6. Include monthly totals as denominator to convert total sales to % of total
7. Update plot accordingly

Beautify your plot

8. Include month labels rather than #'s (Jan, Feb, Mar..). There's a long way to do this and a short way!
9. Reorder months if needed.
10. Cap the y-axis at 6%, set breaks at every 1%, and display as %'s rather than decimals (`scales` package for this).
11. Add a horizontal line at our goal of 3%. Color it and make it dashed.
12. Color the bars based on whether or not you hit the 3% mark.
13. Add a vertical line at November to emphasize before-and-after you joined the team (this is probably too arrogant to do in real life, but we're learning about `geom`'s here!).
14. Finally, try some themes until you find one you like! (`ggthemes` package)

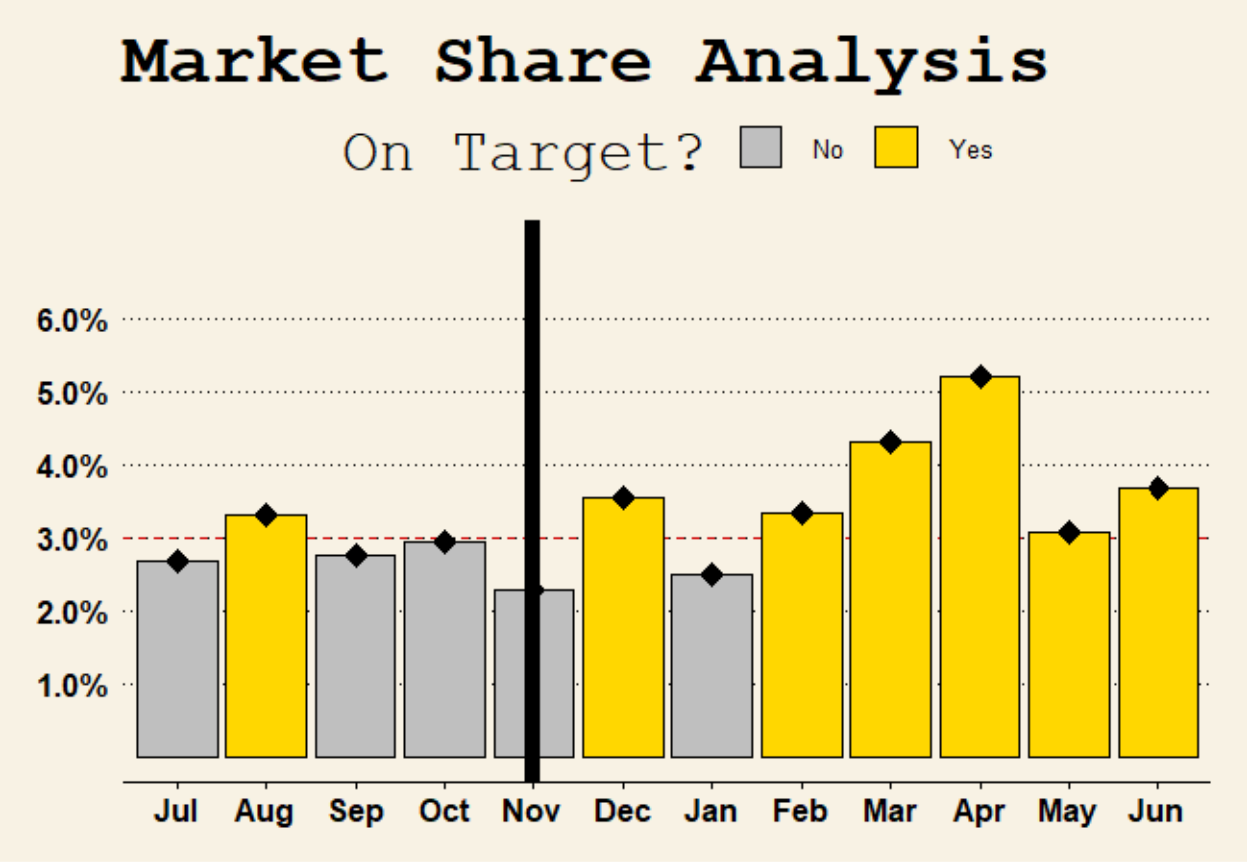
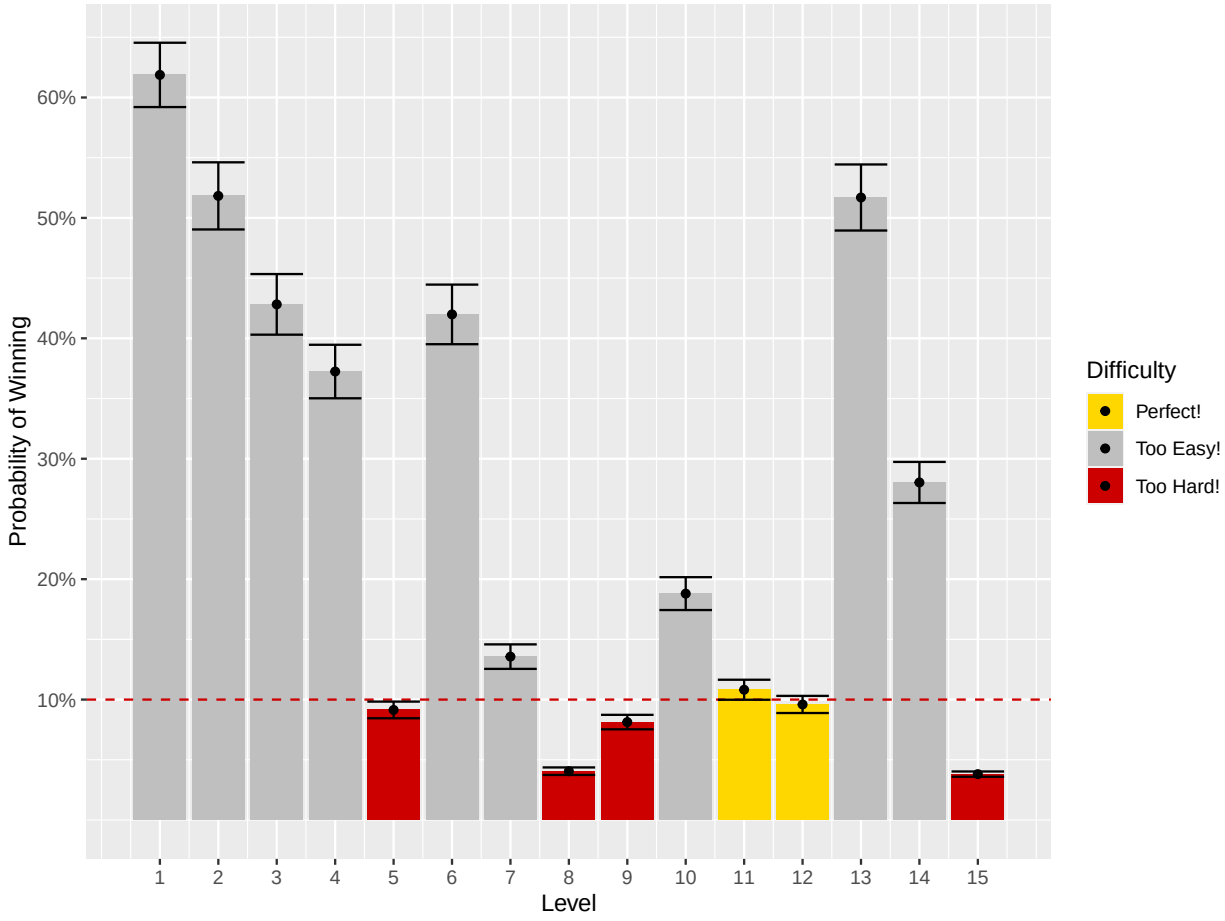


Figure 1: Market Share Analysis

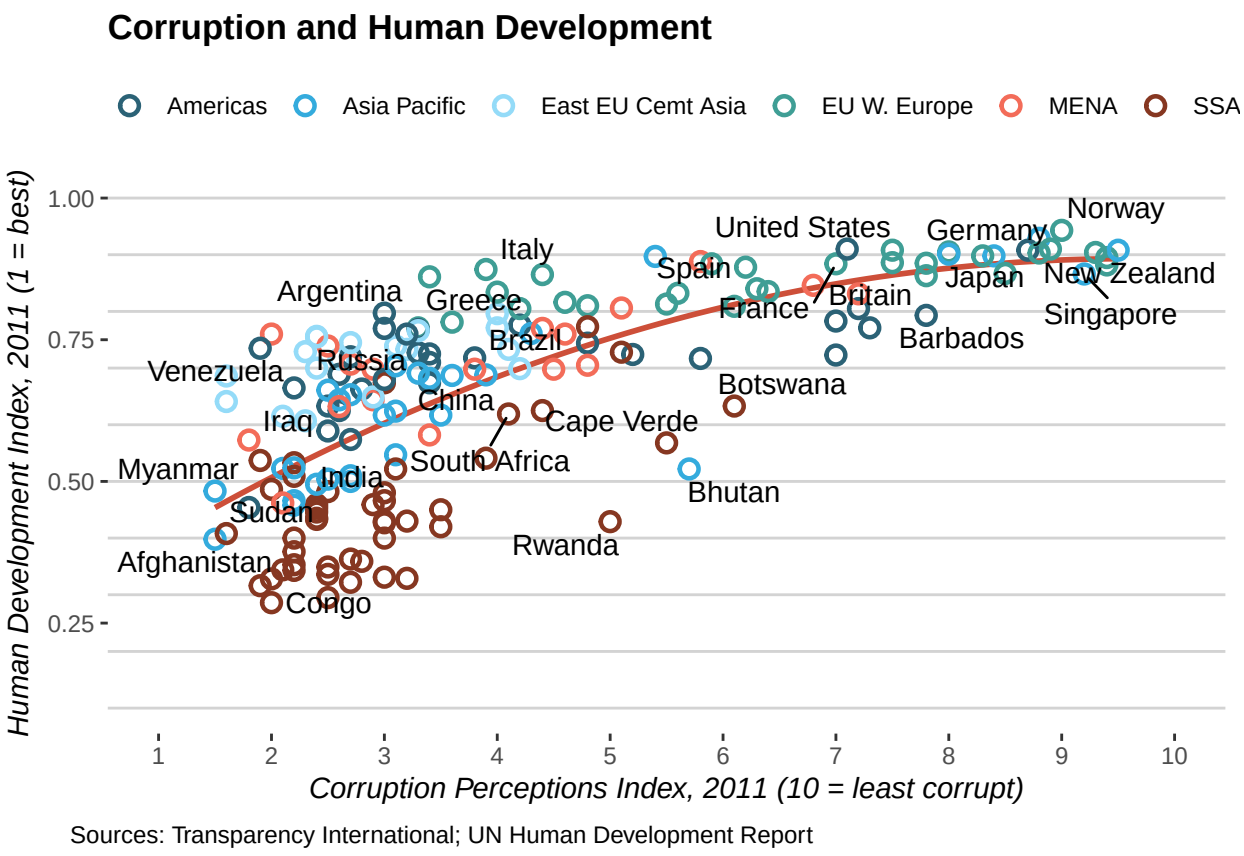
Game Level Difficulty

1. Understand the data
    - What does it contain?
    - What does each row represent?
    - How many players are represented? Is this a lot? Why does that matter?
2. Using `dplyr`, calculate the "difficulty" of each level.
    - Define this as the win percentage
    - Group by level.
    - Calculate total wins and total attempts.
    - Use those to calculate win percentage, p
3. Plot this data using `ggplot`.
    - Level on the x-axis.
    - Win percentage on y-axis.
    - Column (bar) plot
    - Format x axis breaks
    - Format y axis as percentages (**SOLUTION:** using `scales` package, specify `labels=percentage` inside appropriate `scale_...` function).
4. Identify levels that are "too difficult" (less than 10% win percentage)
    - Excessively difficult levels may cause players to get frustrated and quit.
    - Draw a horizontal, red, dashed-line at 10%.
    - Should the level designer be concerned?
5. Incorporate uncertainty.
    - We have *estimates*, but if we want to draw inferences we need *errors* (for intervals, p-values, etc...)
    - We took a sample of players, if we find a 9% win probability for a level, that does not mean the *true* win probability is under 10%.
    - Go back to step 2, and add another column to the data: `error`, calculated as $\sqrt{p(1-p)/attempts}$.
6. Visualize the uncertainty.
    - Now we have estimates and a standard error, we can incorporate confidence intervals for better decision making.
    - Add a `geom_errorbar` layer.
    - Inside the `aes()`, pass the aesthetics `ymin=` and `ymax=`.
        - Set these equal to the appropriate values for a 95% confidence interval.
7. Questions:
    - What hypothesis are you testing when you look at these confidence intervals?
    - Should the level designer be concerned?
    - Which levels would you recommend they modify, if any?
    - Suppose there's also concern about the game being too easy.


Probability of Winning
60%
50%
40%
30%
20%
10%
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Level
Difficulty
Perfect!
Too Easy!
Too Hard!

HDI vs CPI from Economist

1. Create a scatterplot with CPI on the x-axis and HDI on the y-axis.
2. Color by region.
3. Change the shape of the points and increase the size.
4. Change the colors used.
5. Add a loess smoothing line, change the color, remove the error bands, and increase the smoothness via the span option.
6. Change the x and y-axis labels. **BONUS**: Make the italic.
7. Add a title. **BONUS**: Make it bold.
8. Add the caption. **BONUS**: Move the caption to the appropriate position.
9. Move the legend to the top.
10. Force the legend to one row.
11. Remove the legend title.
12. Change the plot background to white / no color.
13. Add the horizontal tick marks.
14. Fix the x-axis breaks.
15. Add country labels by defining a label aesthetic and adding a `geom_text()` layer.
16. Take a subset of countries. Use this subset inside `geom_text()` by redefining the `data=` and `aes()` inside `geom_text()`.
17. Try `geom_text_repel()` instead of `geom_text()`. You need to load the `ggrepel` package for this.



Sources: Transparency International; UN Human Development Report

ORIGINAL (FROM ECONOMIST):

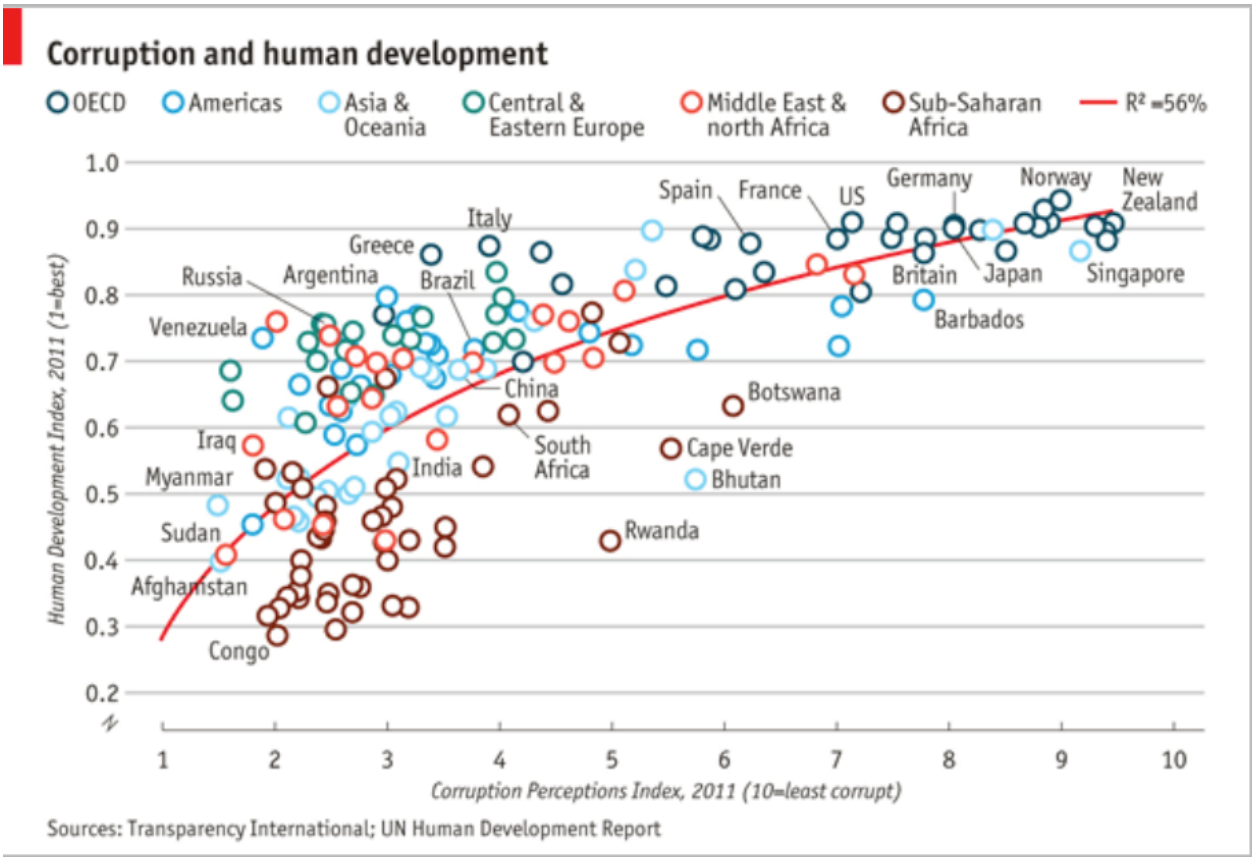


Figure 2: HDI vs CPI

Crime in Florida

1. Read in `jail` dataset, `separate` for latitude and longitude.
2. Create a variable that indicates if a crime was a murder.
3. Filter for rows in Florida
4. Use `map_data()` to get polygon for Florida - `map_data('state', 'FL')`
5. Change the background color.
6. Add crimes to map.
7. Use the indicator for murder that you created to color points.
8. Change colors - red for murder, black for everything else.
9. Play with plot options to improve display.
10. Load `ggthemes` package, try some themes.

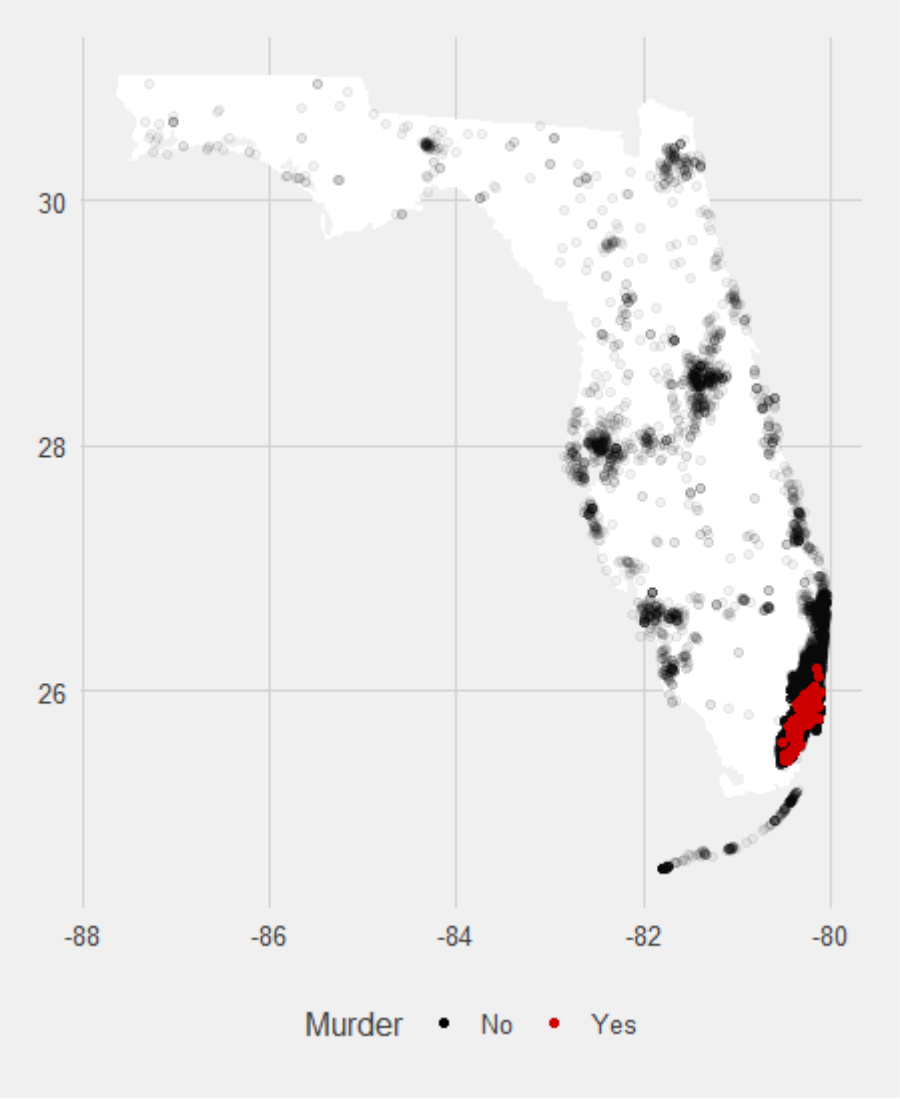


Figure 3: Crime in Florida