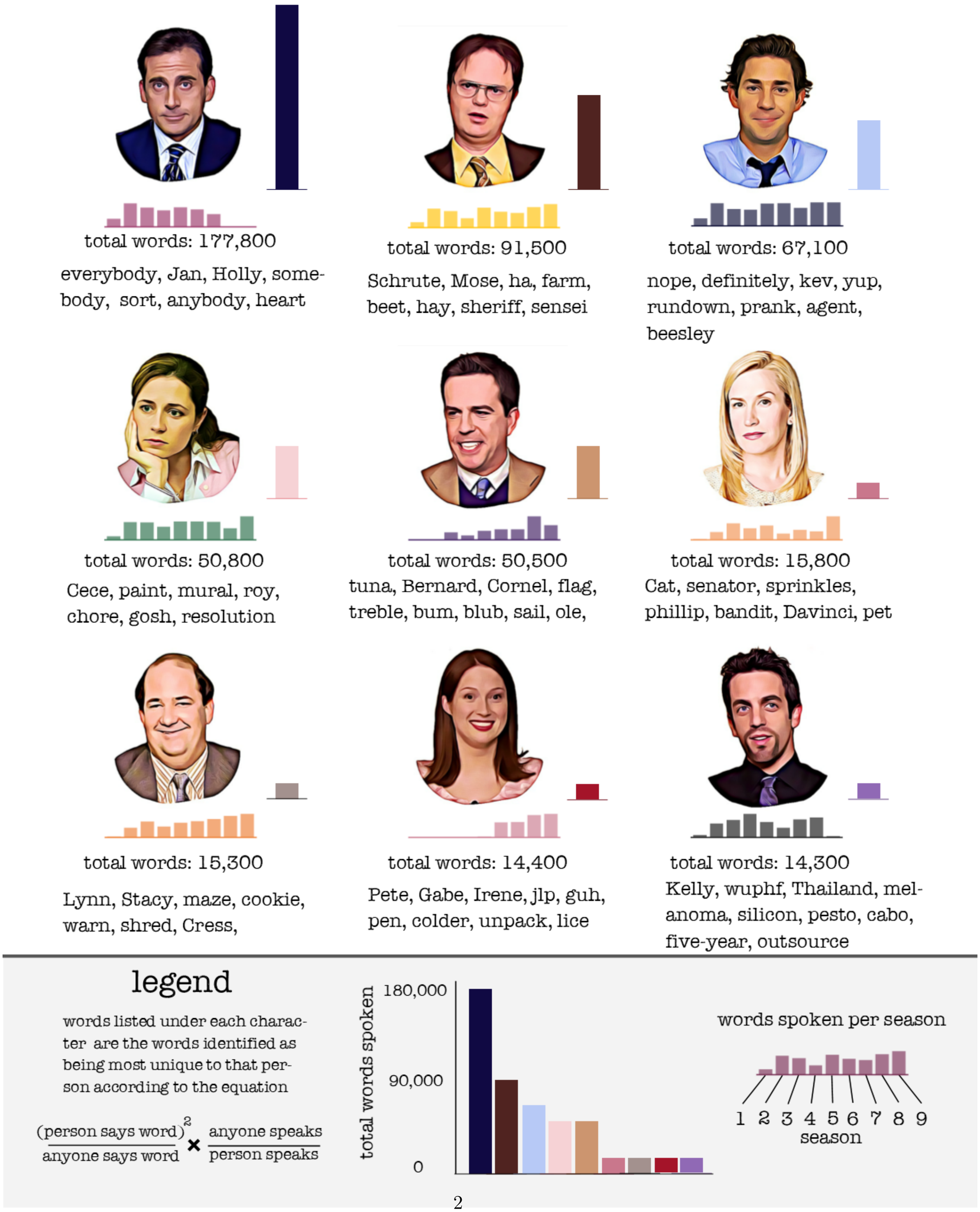
Visualization Critique

Jessica Jiang

4/4/2018

Part 1: Visualization critique



total words: 177,800

everybody, Jan, Holly, somebody, sort, anybody, heart

total words: 91,500

Schrute, Mose, ha, farm, beet, hay, sheriff, sensei

total words: 67,100

nope, definitely, kev, yup, rundown, prank, agent, beesley

total words: 50,800

Cece, paint, mural, roy, chore, gosh, resolution

total words: 50,500

tuna, Bernard, Cornel, flag, treble, bum, blub, sail, ole,

total words: 15,800

Cat, senator, sprinkles, phillip, bandit, Davinci, pet

total words: 15,300

Lynn, Stacy, maze, cookie, warn, shred, Cress,

total words: 14,400

Pete, Gabe, Irene, jlp, guh, pen, colder, unpack, lice

total words: 14,300

Kelly, wuphf, Thailand, melanoma, silicon, pesto, cabo, five-year, outsource

legend

words listed under each character are the words identified as being most unique to that person according to the equation

$$\frac{(\text{person says word})^2}{\text{anyone says word}} \times \frac{\text{anyone speaks}}{\text{person speaks}}$$

URL:https://i.redd.it/3k2o82q966q01.png

This visualization describes in The Office, the total number of words in 7 seasons as well as a number of words spoken by each main character in each season separately. It also lists the unique words that the character speaks, which can be used to identify different characters. At the bottom of the visualization, there are four different legends: an explanation of the words listed under each character, the equation used to calculate the frequency of words and identify the unique words, a histogram of total words spoke by characters, and a sample histogram used to demonstrate that every bar chart stands for amount of words spoken by that character. In the following paragraphs, I will analyze this visualization based on the five qualities of great visualizations: truthful, functional, beautiful, insightful, and enlightening.

First, from the perspective of being truthful, assuming the data provided by April's DataViz competition is reliable, the author still failed to provide a much more accurate and factual depiction of the topic of 'The Office' Characters' Most Distinguishing Words. Instead of visualizing complete information of all 29 seasons of scripts, he only depicts data from 7 seasons. We do not know which 7 seasons they are and if these 7 seasons are representative of the show and thus not sure if information summarized from these 7 seasons can be useful for readers to generalize the characters through 20 seasons. If the author would like to focus on these 7 seasons, he should have specified in the caption or in the subtitle with more details about the 7 seasons. In addition, the equation used to calculate the total words is not weighted by the timing of when the character first appears. For example, if a character just starts appearing in the 7th season but becomes actually much more popular in the latter shows, the unique words identified by this equation would not be representative of his language. A vocabulary that appears only in one or two episodes, or first appears in the last episodes for a character that appears in 7 seasons, for example, does not represent a word uniquely associated with the character. The bar chart at the bottom of the visualization of total words spoken is ambiguous. I would recommend the author to add a filter and to only consider those words that appear in at least 3 or 4 episodes.

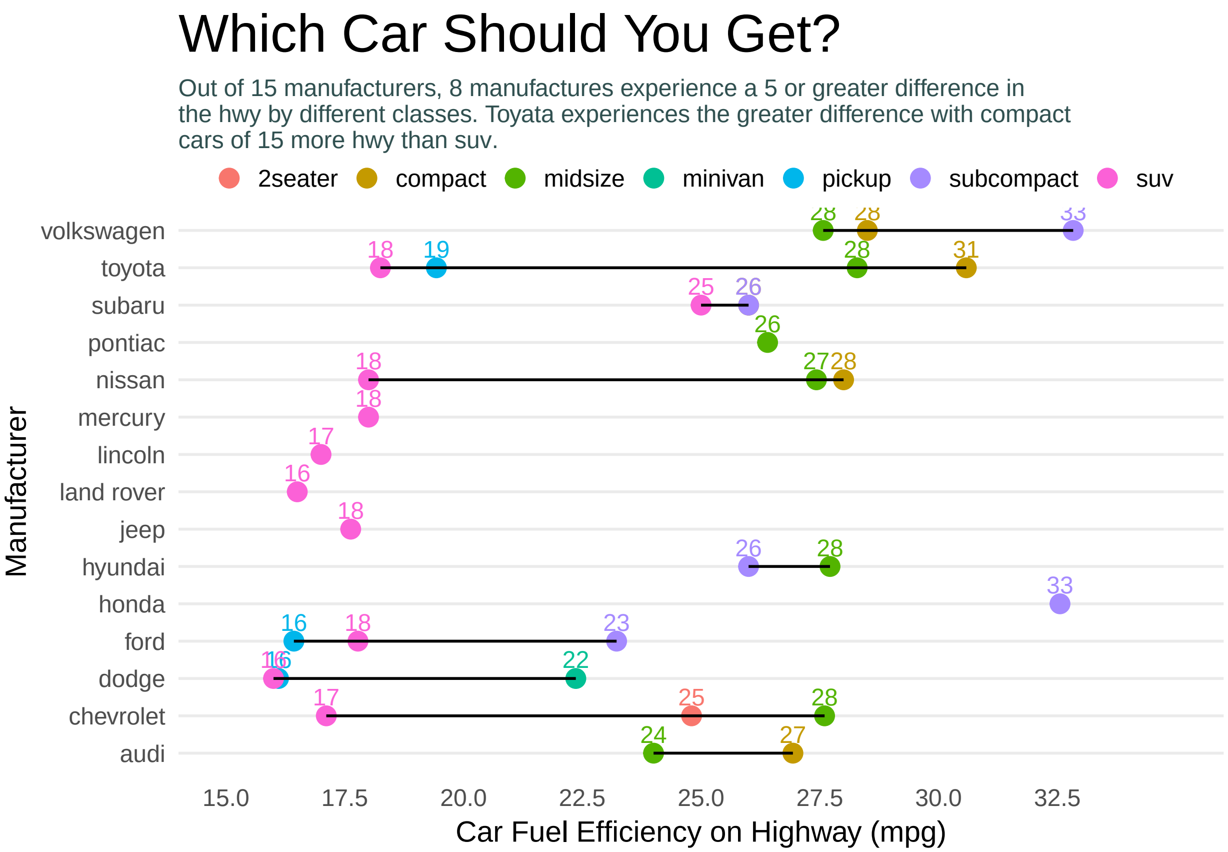
Second, while the image is considered functional for readers to get an idea of the unique words that might represent a certain character, the scale is ambiguous in the bar graphs under each character. Other than a general trend of the total amount of words, there is no way where we can figure out how much exactly the person speaks in a season. Meanwhile, readers might not be able to compare within each character the frequency of the unique word. For example, we don't know either "every day" or "Holly" appears more often for Michael. It would be great if the author can give more numeric details.

This visualization is beautiful in the sense that it increases its novelty and visual appeal by adding a portrait of each character, which is entertaining to look through and enables readers to identify the character more easily. Each character is portrayed in the way that his/her emotion and personality are expressed through facial expressions and postures. It is effective when the results actually paint a picture of the character and when readers can link unique words with their appearance. Other than that, However, this visualization only gets the job done, but not very efficient at mixing sensual and intellectual pleasure. By using two different colors for each of the characters does not aid readability and efficiency at letting readers extract patterns and make accurate comparisons. Meanwhile, with the relatively big portraits, there is not enough space for more comprehensive information of each bar chart and word. Readers are not able to extract more detailed information as well as the trend among characters. In a word, beauty is a very important attribute for readers to get attracted and to decided if they want to explore more from the visualization. However, designers should not compromise functionality to just make prettier graphs. A balanced combination of beauty and functionality is important.

I would not consider this visualization insightful because the information from this visualization is sort of trivial and worthless for most of the readers. It does not generate enough insights that can be used to build knowledge. For people who are big fans of the show, it might be a bit insightful for them to know the words that can be used uniquely to identify the characters, and then? Would readers really use the information to identify characters in the show or even to analyze characters' personality? The answer is more likely to be negative. Besides, there is not any interaction visualization and does not get readers to engage. As the result, the visualization loses the chance to provide readers with a process of exploring deeper information.

Finally, this is not an enlightening visualization. For most readers, the topic of 'The Office' Characters' Most Distinguishing Words does not ethically relate to any important or enlightening issues and is not critical to the well-being of most people. There is no further delighted that the readers can obtain the density of data they can explore. Since this topic is given by the competition, we do not want to blame it on the author. The author did a good job at providing a fun and entertaining visualization, but he could have done better if more numeric details and scale adjustments could be added to the visualization.

Part 2: ggplot2 and the grammar of graphics



With the mpg data frame found in ggplot2, I want to depict the relationship between fuel efficiency on the highway among groups as well as the relationship among different classes within each manufacturer. For example, whether the compact VW on average is more fuel efficient than a compact Nissan, and how much more miles per gallon Ford minivan can reach compared to a Ford pickup.

There are three important variables in this visualization: `car manufacturer`, `why` and `class`; `car manufacturer` stands for the manufacturer of a car, `hwy` is a car's fuel efficiency on the highway in miles per gallon, and class is the classification of the car. In order to visualize the relationship between these three variables in the data set, instead of plotting all data points which can make the plot too crowded and not generalizable, I aggregated the data by manufacturer and class, and computed and plotted the mean of fuel efficiency for each group. Since we want to include three dimensions in one visualization, multiple types of information need to be combined and compared. To get a general sense of the distribution, I first tried a dodged barplot (**Exhibit 1**). It looks just BAD. Too many things are happening there: length of bars is too long, some space is wasted, labels are overlapped on an x-axis, comparisons are not clear, and etc. Then I decided to use a stacked bar chart (**Exhibit 2**) to combines some bars for different classes by the same manufacturer. The visualization looks better here; and after I flipped the coordinates (**Exhibit 3**), labels are not overlapped anymore. However, the visualization still does not provide a good sense of how cars of different classes differentiate. Besides, in this case, where the difference is relatively small here, the comparisons become visually complex.

In order to reduce the clutter and focus on the mpg by each class, I decided to use scatter plot and to map variable class to color aesthetic (**Exhibit 4**). It looks much better and prettier now: while multiple categories are combined in a compact space, the graph is relatively clean and concise. Besides, I also add the shape aesthetics to the scatter plot (**Exhibit 5**), and it looks a bit messy for me. So I decided to stick with graph 1 without mapping class to shape. However, a problem comes up here: readers still cannot get a fast idea of the fuel efficiency and it could still be hard to make comparisons (mpg of midsize for VW and Hyundai for example). To solve this problem, I add number text of hwy value on the top of each data point as well as a line that connects data points within each group of a manufacturer, which allow us to get a sense of how big the variance of fuel efficiency is within a manufacturer(**Exhibit 6**). The work is not done yet. To have a complete visualization, we need to have a title, caption, legend, and etc. I added these elements to the graph, including a subtitle that has a summary of what story the visualization tells and what I want to emphasize (Graph: **Which Car Should You Get?**). Besides, I also set theme_minimal() and remove the vertical grid that is not that useful here after we add numeric text to each point. Finally, I rearranged the legend of class mapping and displayed them all horizontally and replaced them on the top of the chart; this way we have more space for this relatively wide graph and it is more visually comfortable.

First of all, data is truthfully visualized. All manufacturers are plotted and for each manufacturer, the mean of car fuel efficiency (hwy) by each class is calculated. For manufacturers, Mercury, Jeep, Lincoln and Land Rover, are not removed and hidden from the visualization even they only have one class of SUV. Besides, the scale of the x-axis is set to be the same, which allows us to make the comparisons more easily. Second, it can be considered functional: the difference between classes is more straightforward and clear as I connect the classes for each manufacturer. Also, since sometimes readers really care about the numeric value of the data points, adding value marks to the point can help to clarify the differences between the points, either within or among groups. We can also compare the length of lines to see either how variance mpg is among different classes. Besides, in order to have the plot less scattered, I scaled x-axis by limiting the range as (15, 35) and the break as 2.5. This visualization is relatively insightful and enlightening as it can be used to conduct some further cost-benefit analysis. For a consumer who wants to choose a made of the car and she/he cares about fuel efficiency on the highway, this chart provides a vertical comparison among different manufacturers of the same car class. For a consumer who has already decided the manufacturer, this chart can help to compare the class. I need to admit that if I can add some interactions in this visualization, I believe the data set can be presented in a more compelling way. In summary, I believe visualizing a dataset does not mean we need to display as many data as possible; rather, it means we should craft the data, focus on the story we want to tell, and communicate important aspects to readers. The graph above did a good job of simplistically illustrate and compare your important data points.

Exhibit 1

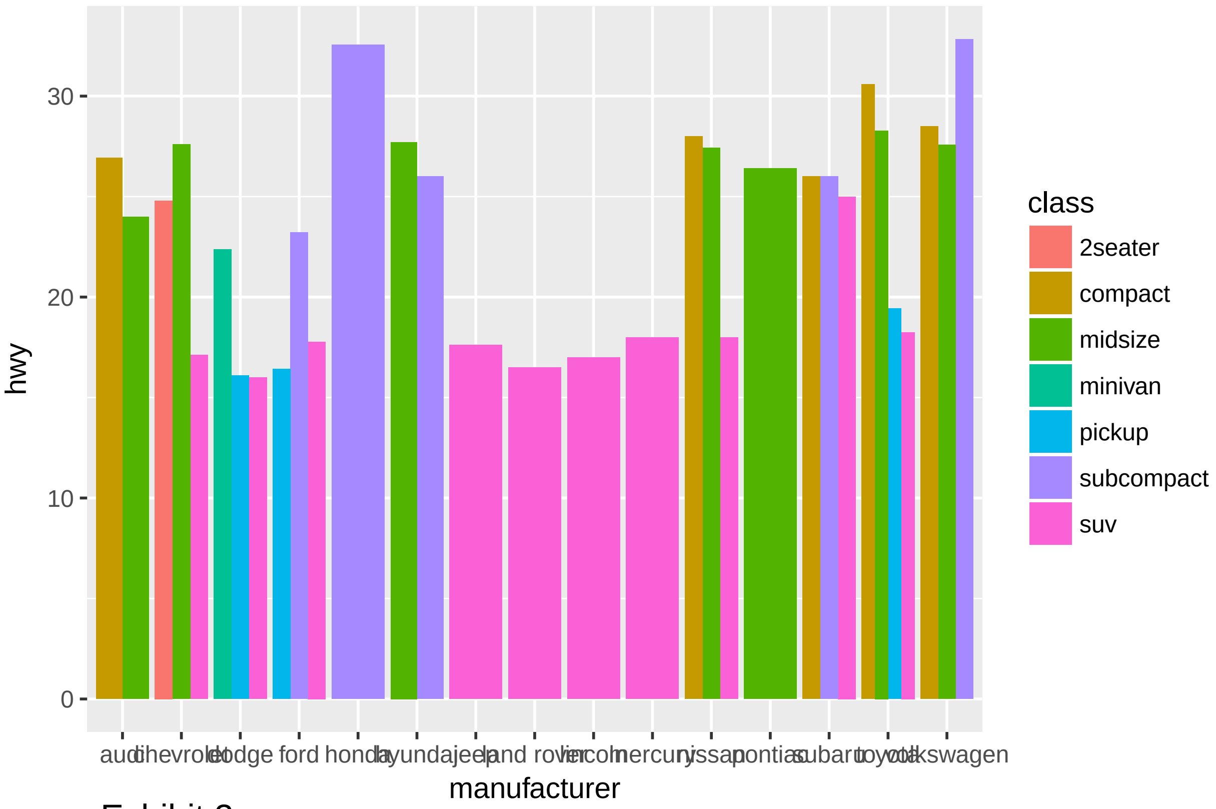


Exhibit 2

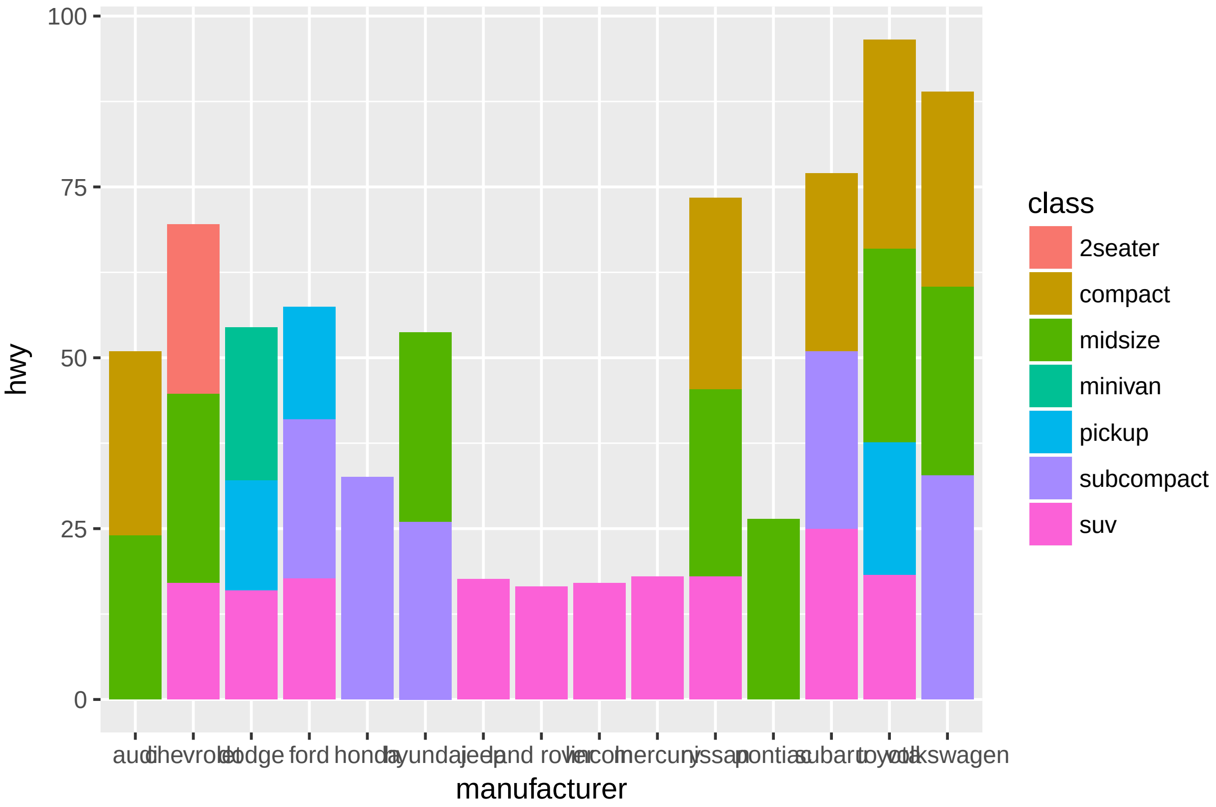


Exhibit 3

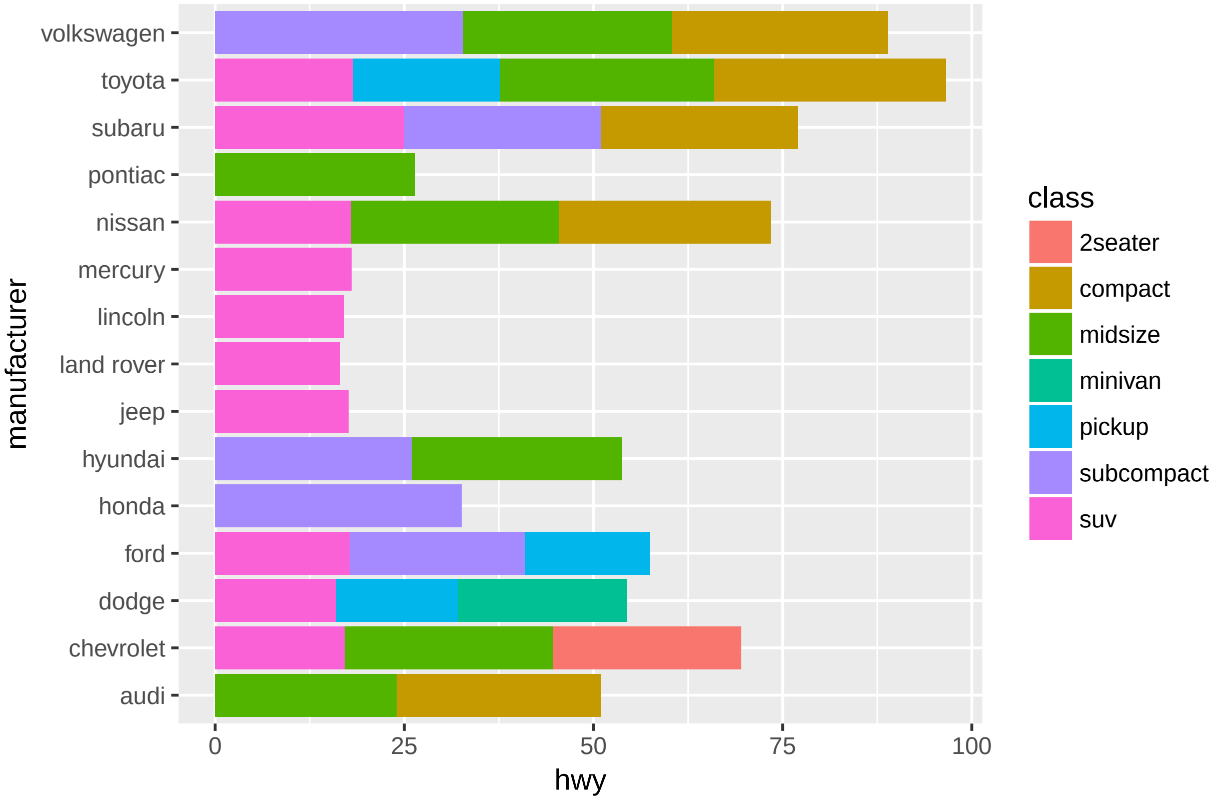


Exhibit 4

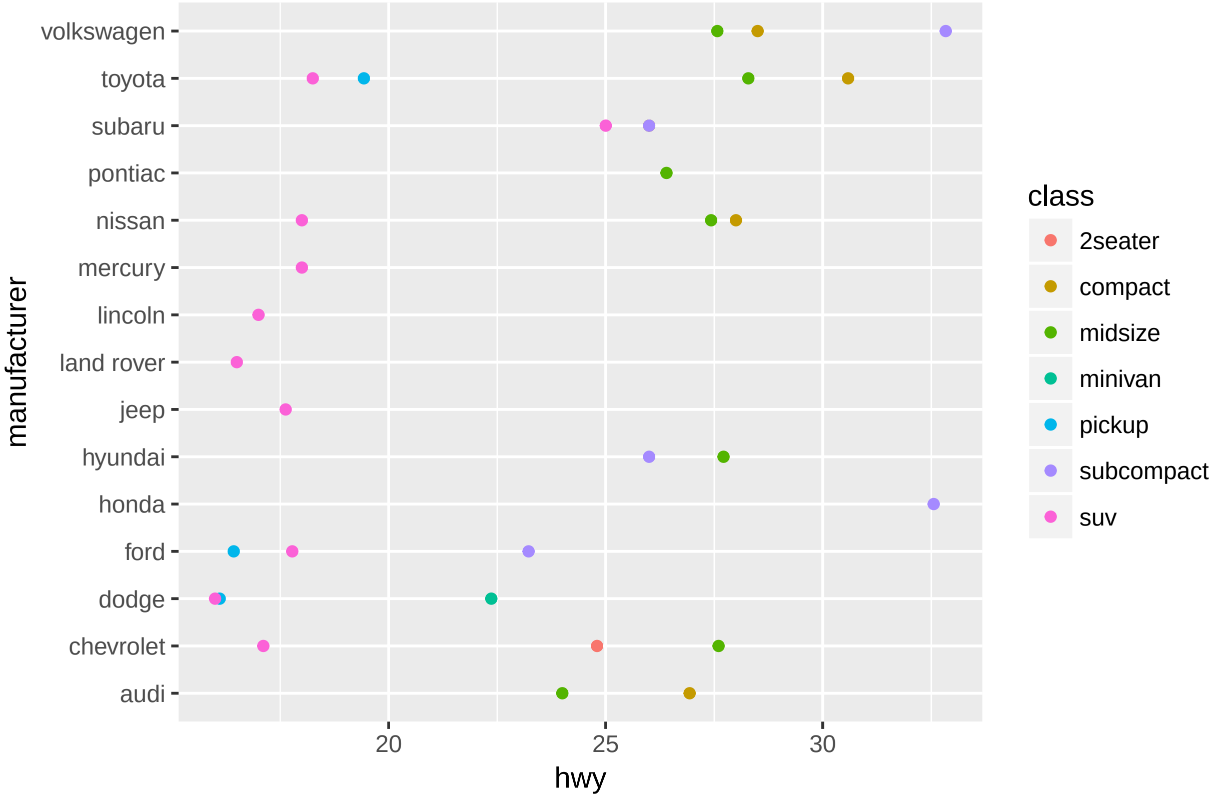


Exhibit 5

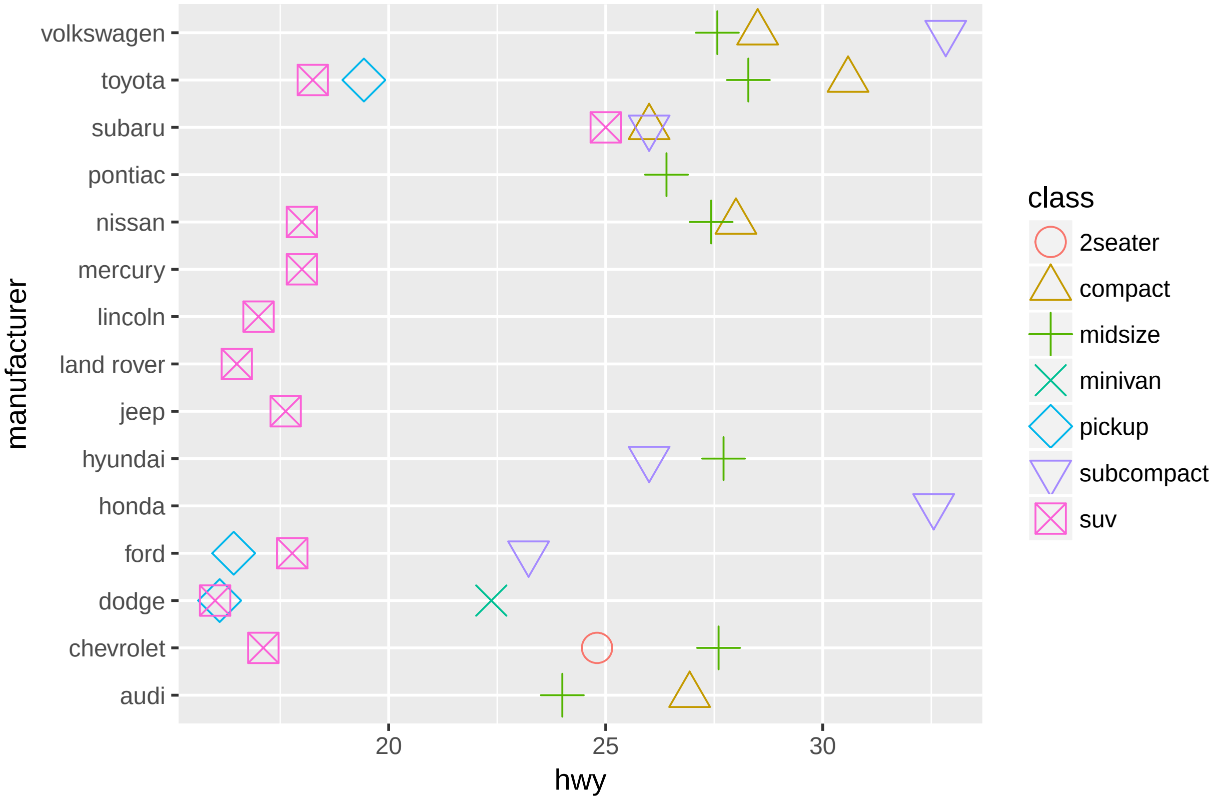
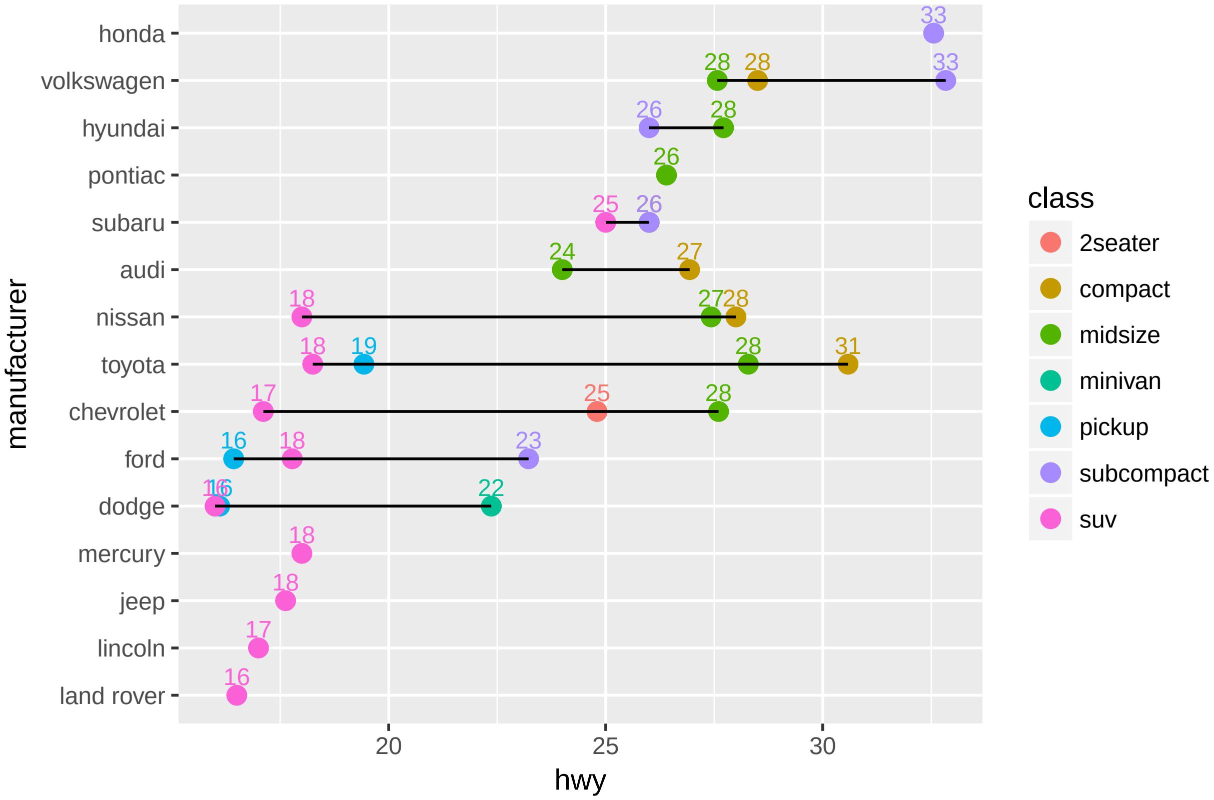


Exhibit 6


manufacturer
honda
volkswagen
hyundai
pontiac
subaru
audi
nissan
toyota
chevrolet
ford
dodge
mercury
jeep
lincoln
land rover
hwy
class
2seater
compact
midsize
minivan
pickup
subcompact
suv