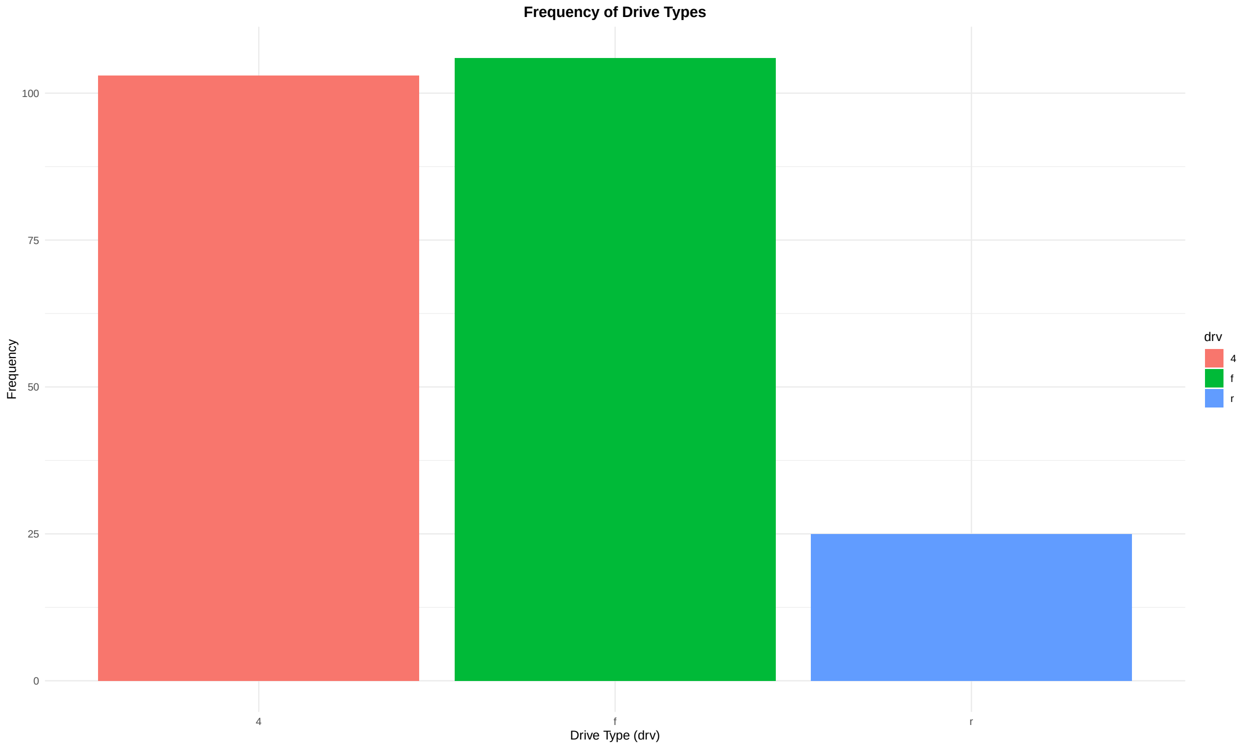
Homework1.Rmd

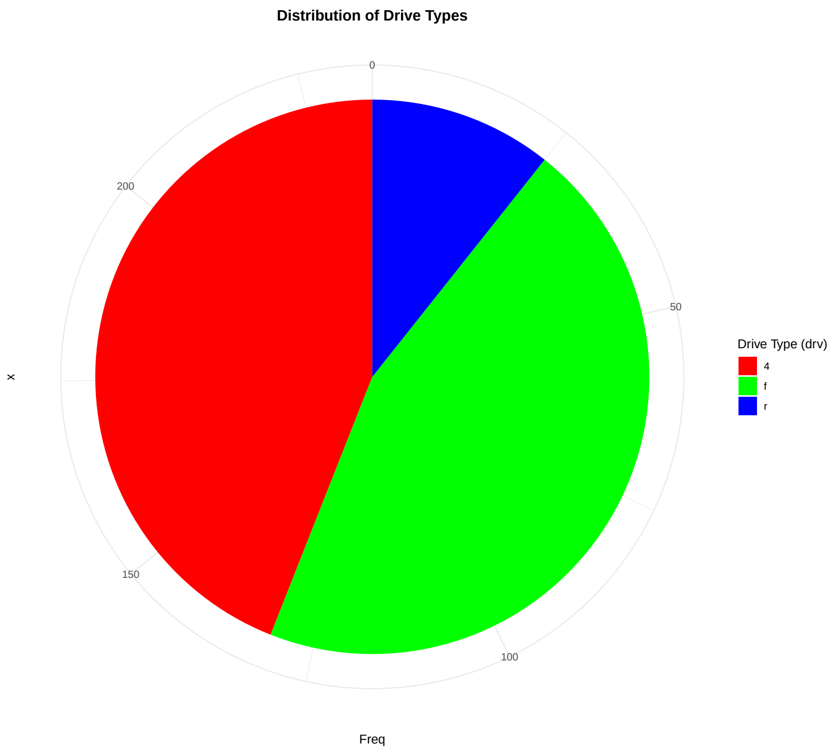
2024-10-10

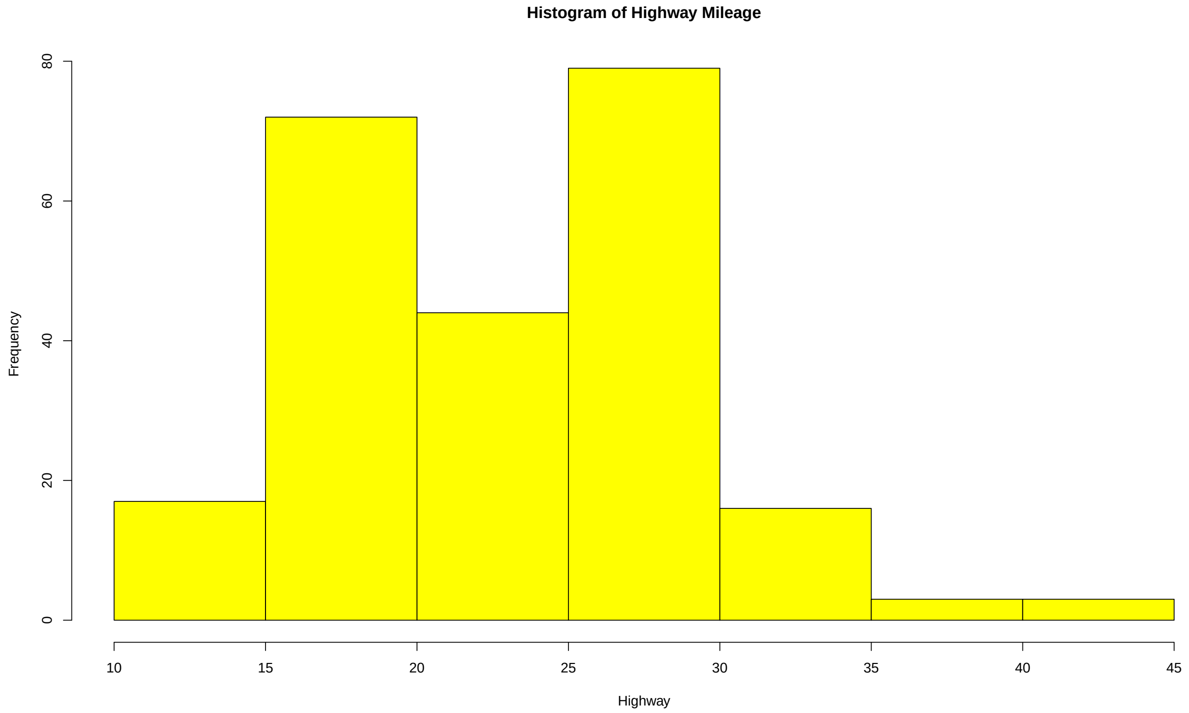
Task 1: Improved Plots

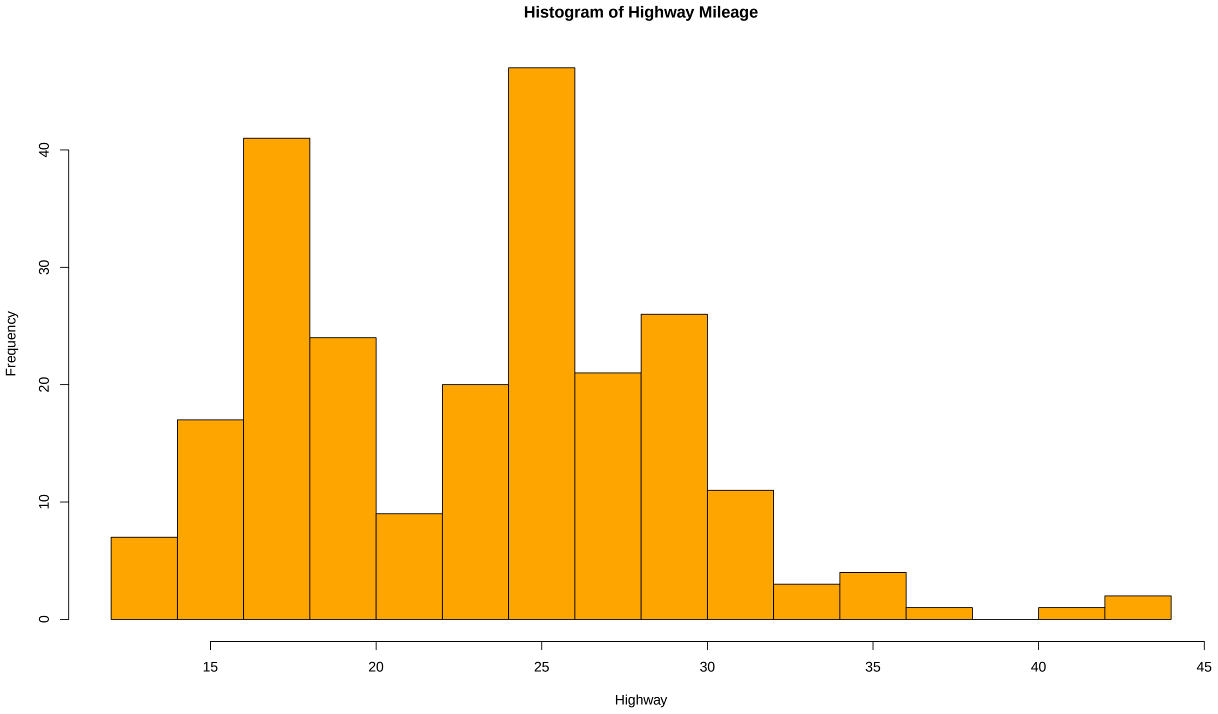
```
##   manufacturer model displ year cyl      trans drv cty hwy fl   class
## 1         audi    a4   1.8 1999   4   auto(l5)   f  18  29  p compact
## 2         audi    a4   1.8 1999   4 manual(m5)   f  21  29  p compact
## 3         audi    a4   2.0 2008   4 manual(m6)   f  20  31  p compact
## 4         audi    a4   2.0 2008   4   auto(av)   f  21  30  p compact
## 5         audi    a4   2.8 1999   6   auto(l5)   f  16  26  p compact
## 6         audi    a4   2.8 1999   6 manual(m5)   f  18  26  p compact
```

```
##   drv Freq rel_Freq Percentage
## 1   4  103     0.44         44
## 2   f  106     0.45         45
## 3   r   25     0.11         11
```

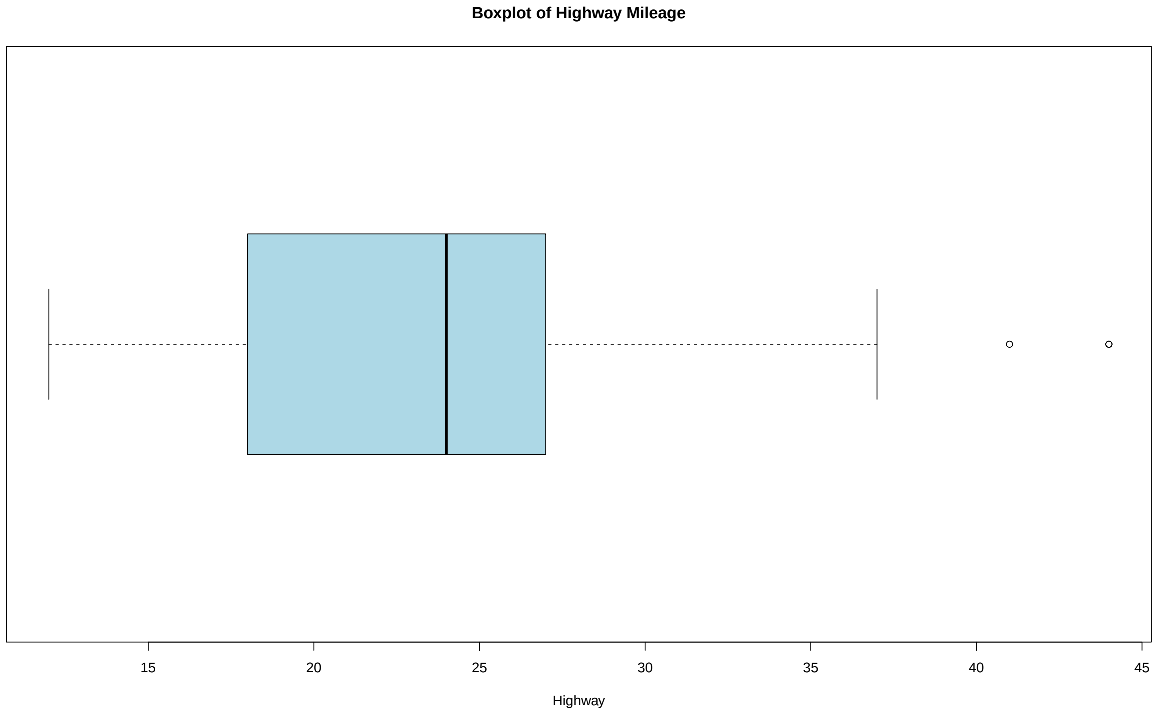



Distribution of Drive Types
0
50
100
150
200
x
Freq
Drive Type (drv)
4
f
r

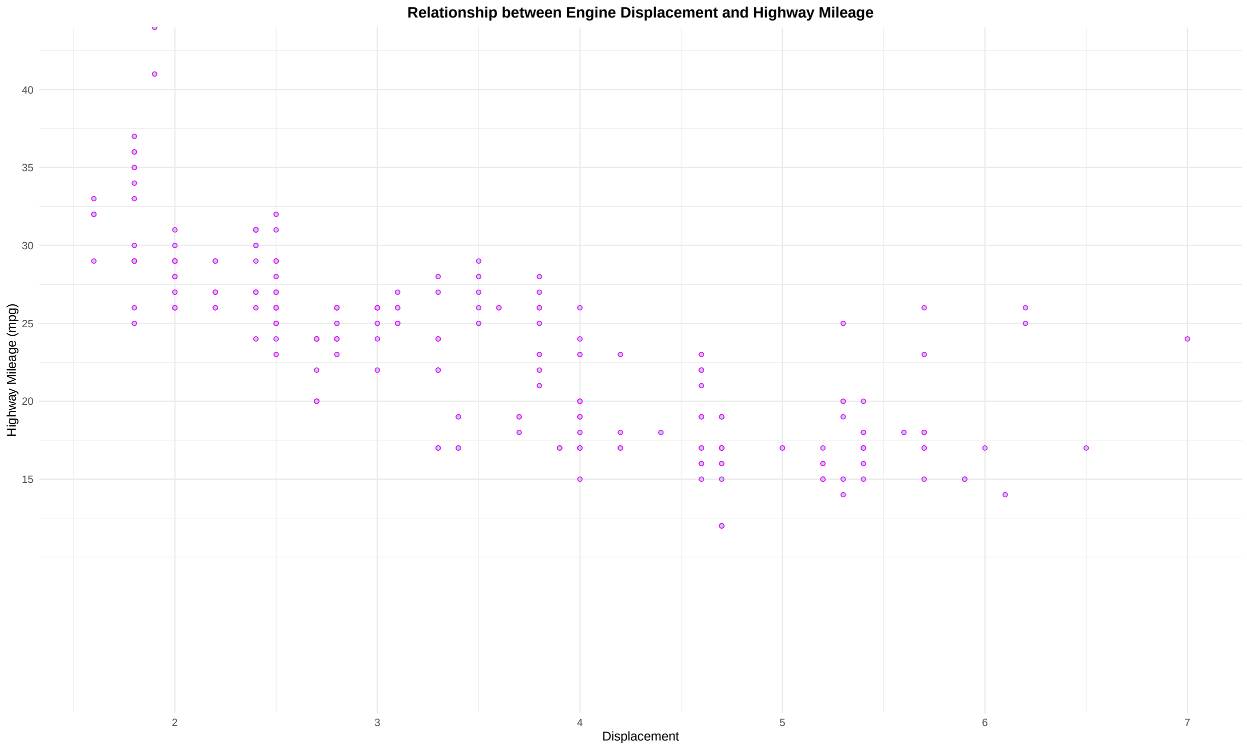

Histogram of Highway Mileage
Frequency
0
20
40
60
80
10
15
20
25
30
35
40
45
Highway

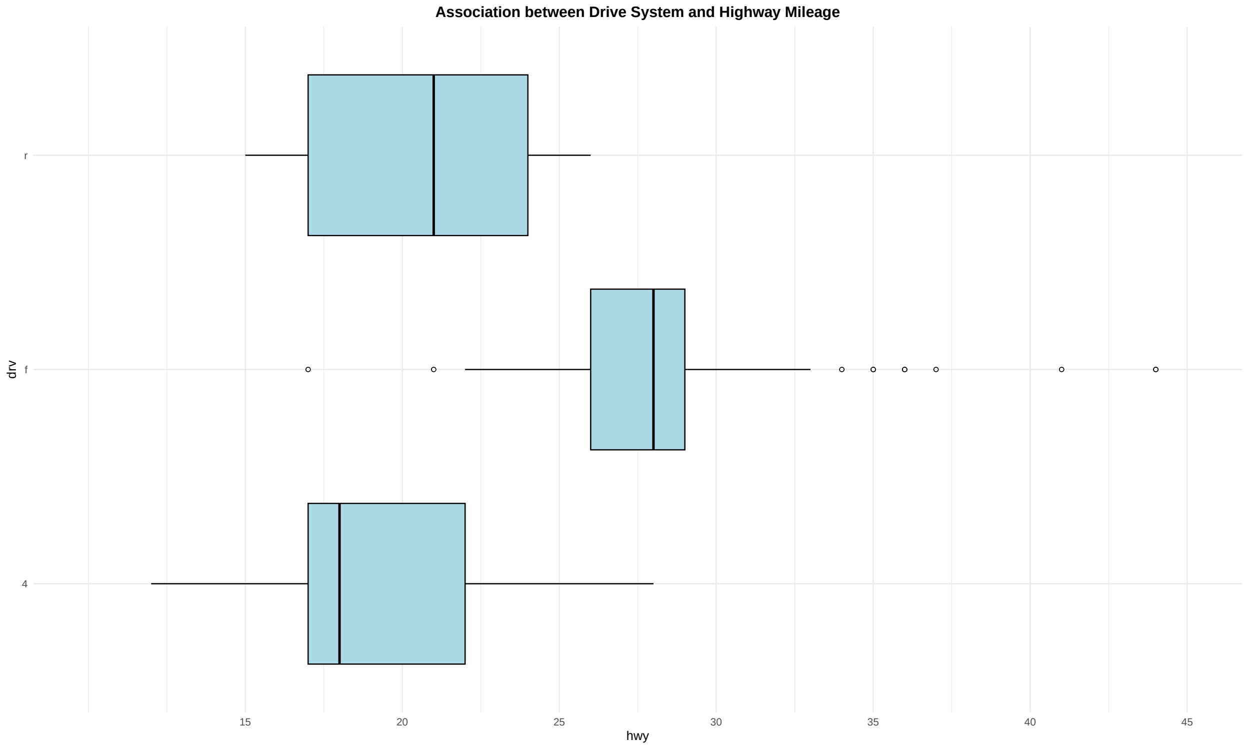


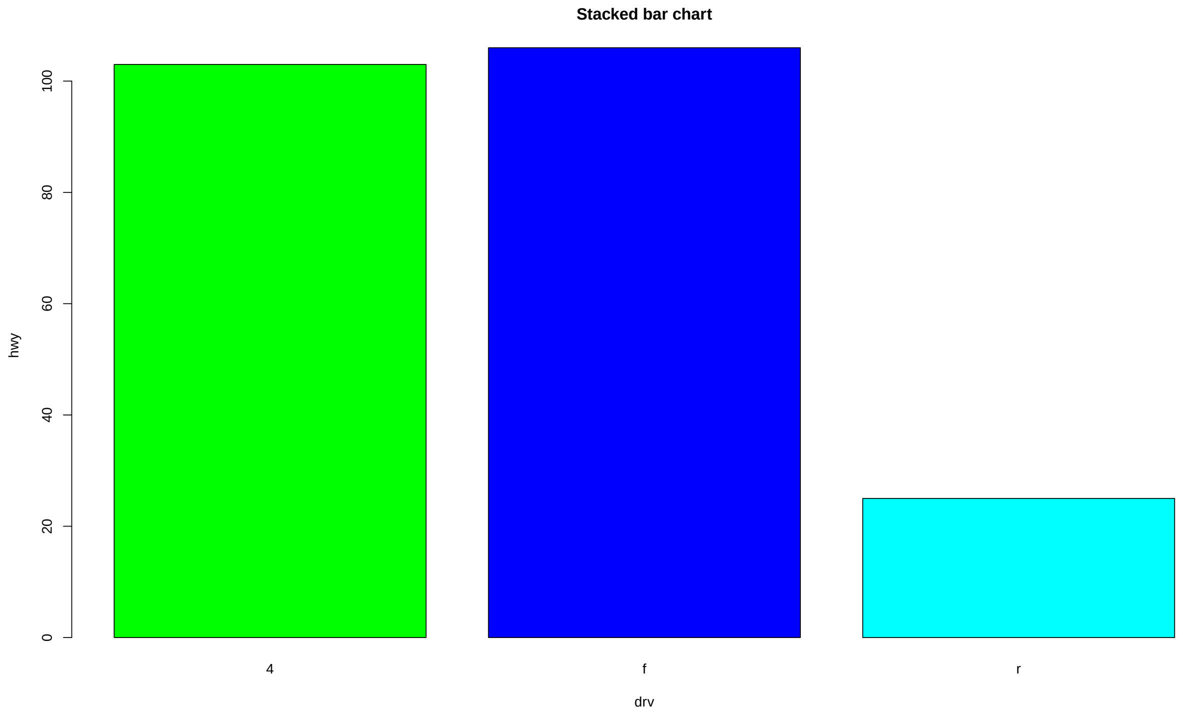
Histogram of Highway Mileage

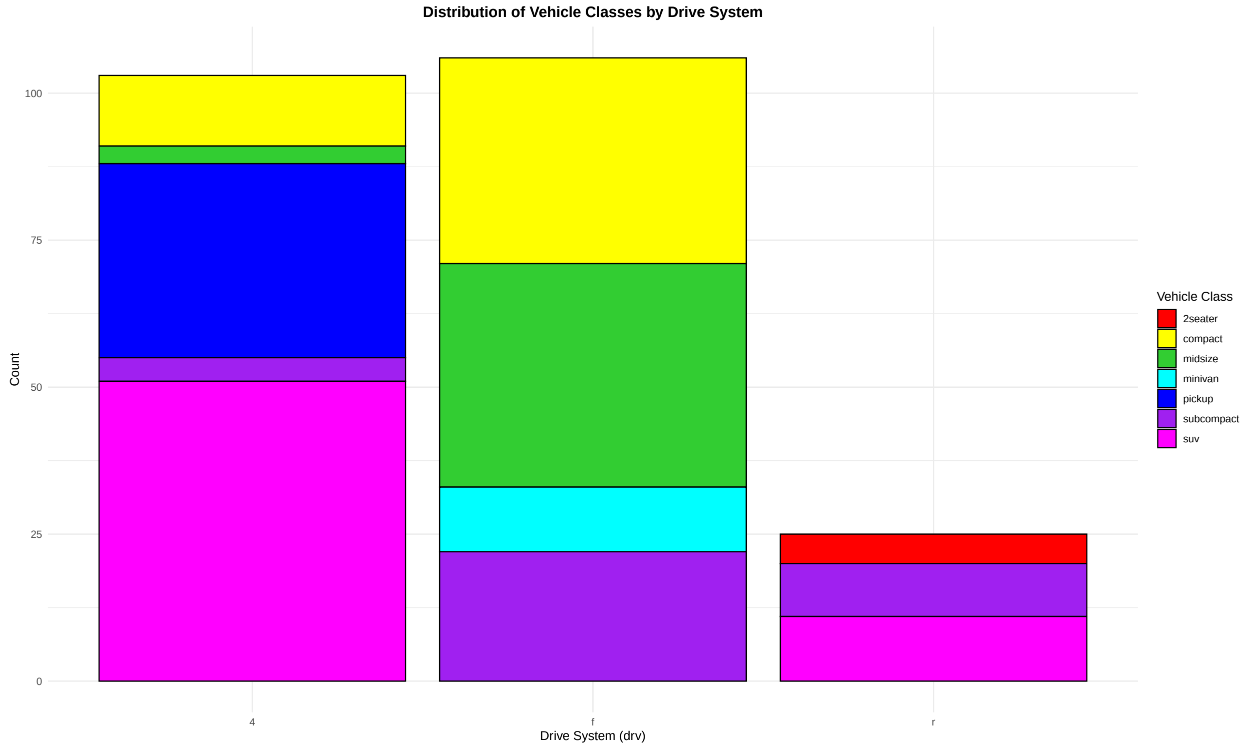


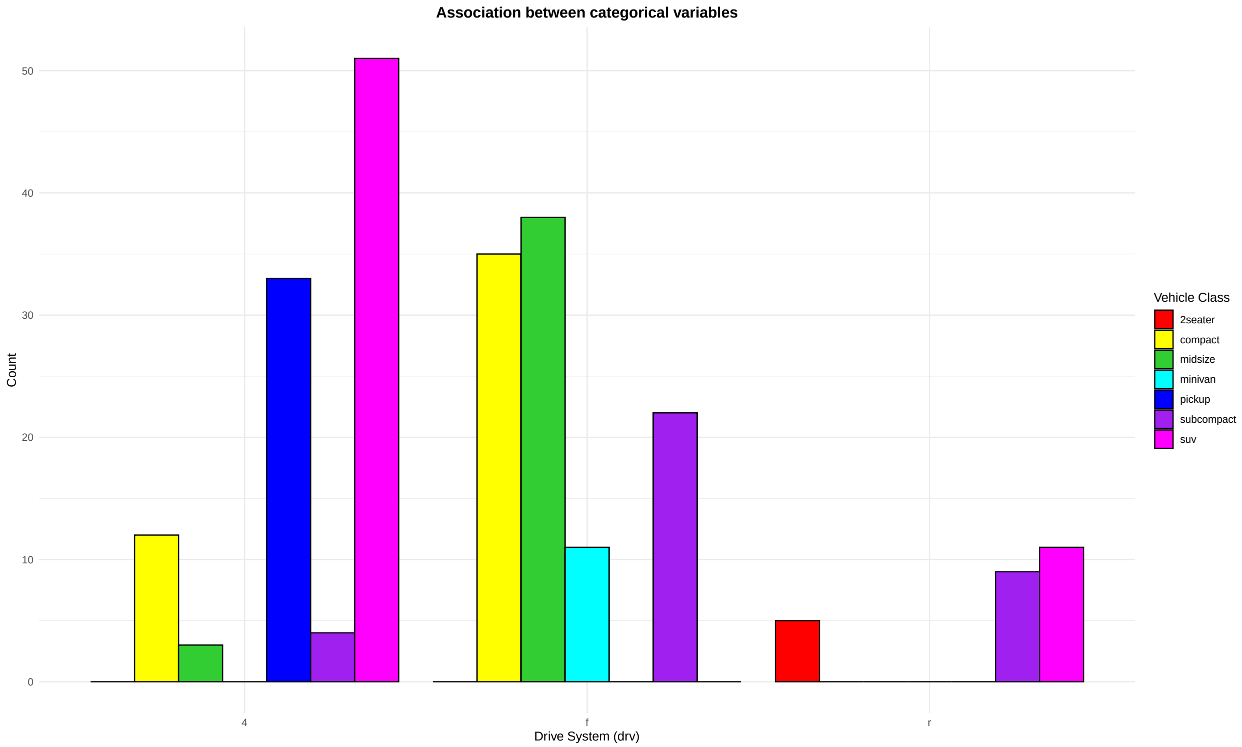
Boxplot of Highway Mileage

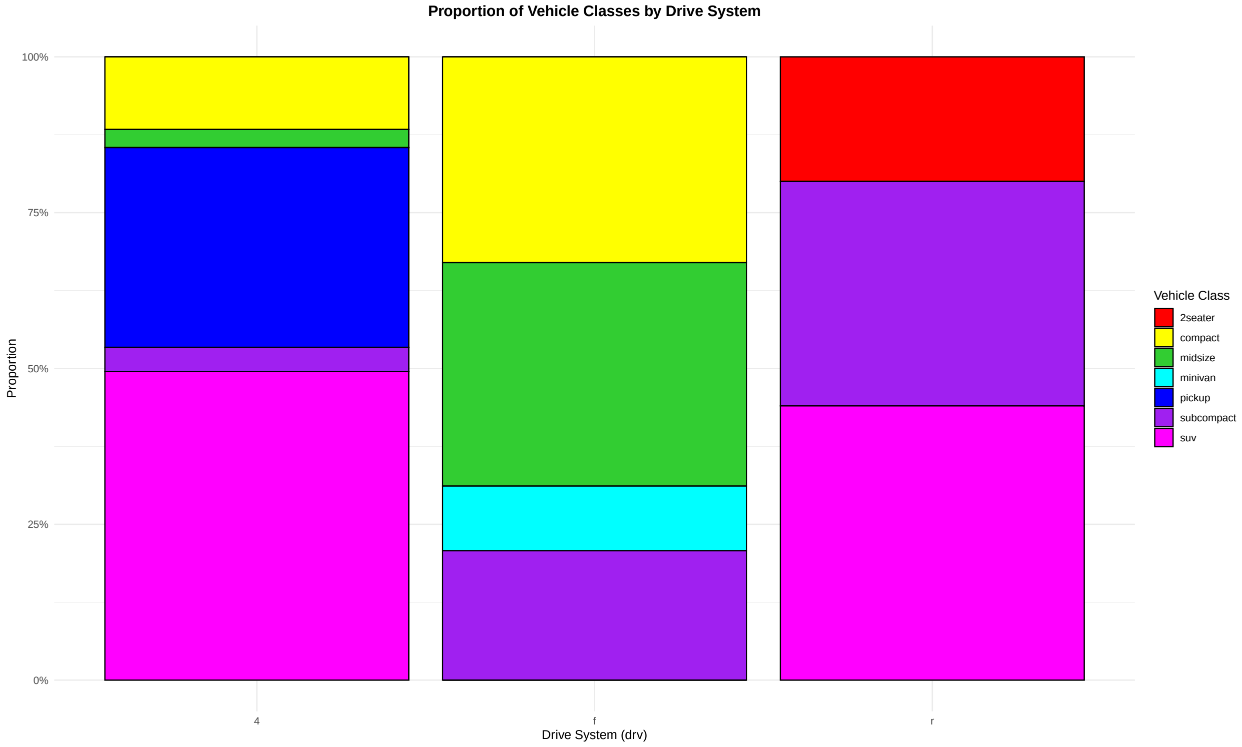

Relationship between Engine Displacement and Highway Mileage
Highway Mileage (mpg)
40
35
30
25
20
15
2
3
4
5
6
7
Displacement

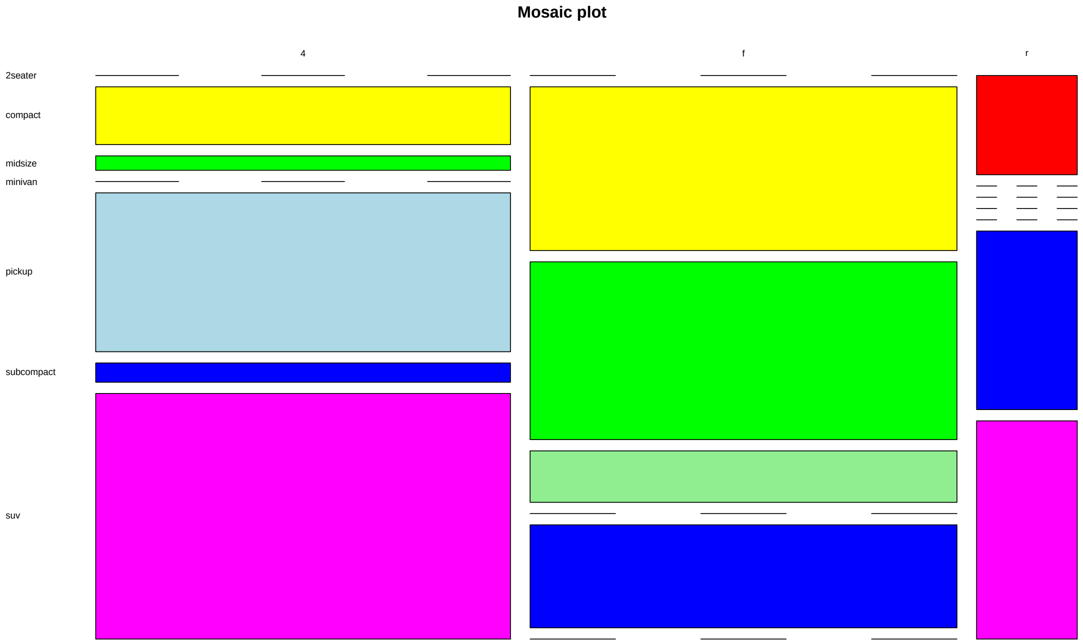

Association between Drive System and Highway Mileage
drv
r
f
4
15
20
25
30
35
40
45
hwy


Stacked bar chart
hwy
0
20
40
60
80
100
4
f
r
drv


Distribution of Vehicle Classes by Drive System
Count
0
25
50
75
100
4
f
r
Drive System (drv)
Vehicle Class
2seater
compact
midsize
minivan
pickup
subcompact
suv

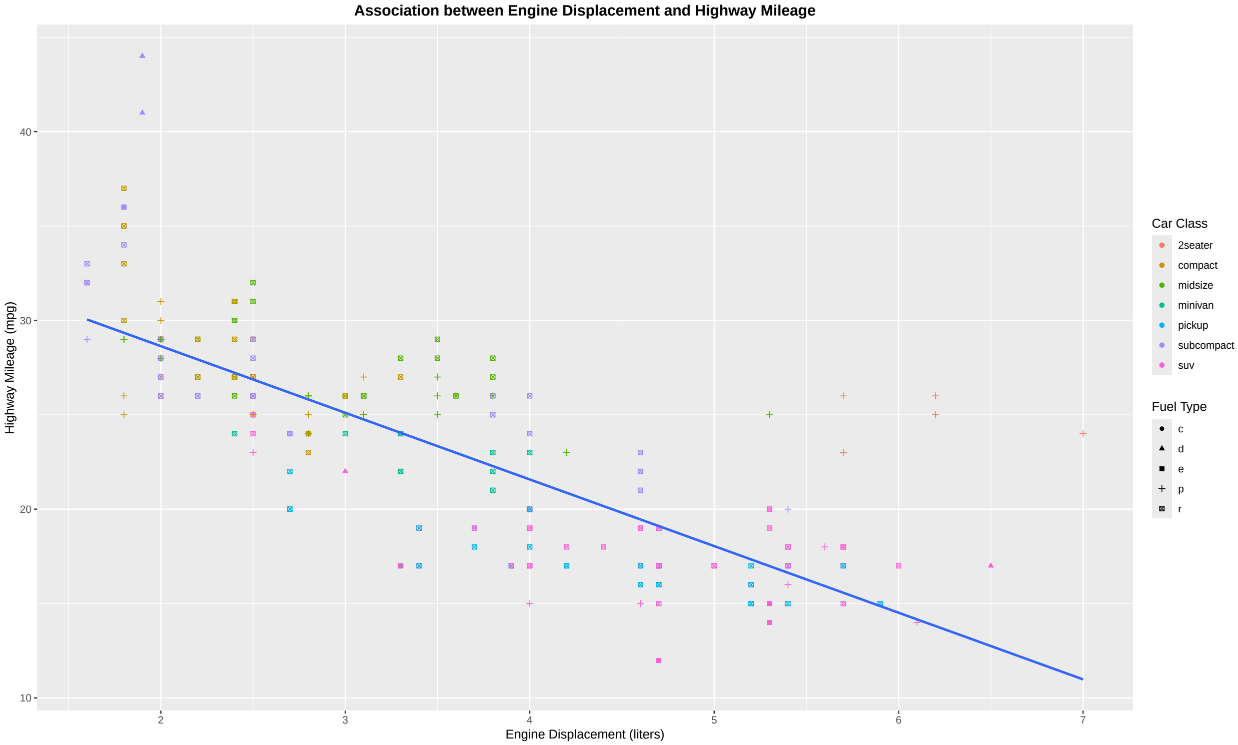

Association between categorical variables


Proportion of Vehicle Classes by Drive System

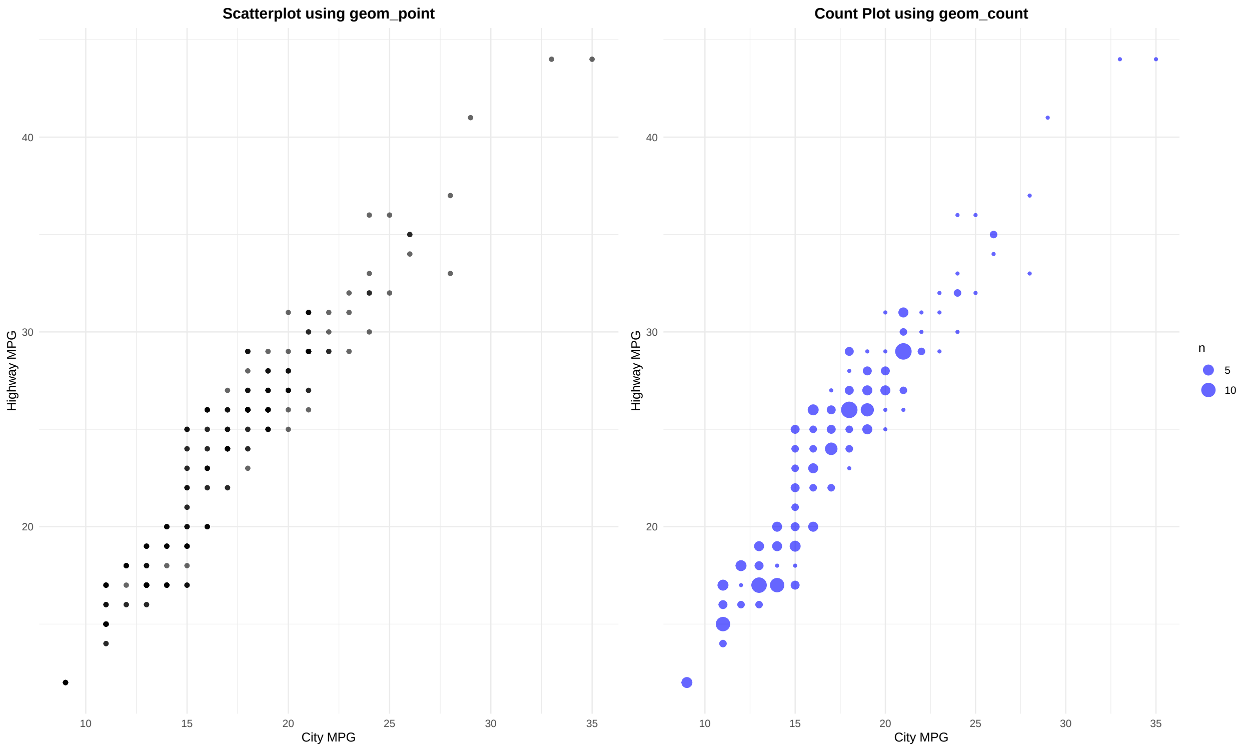


Task 2: Association between engine displacement and highway mileage

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



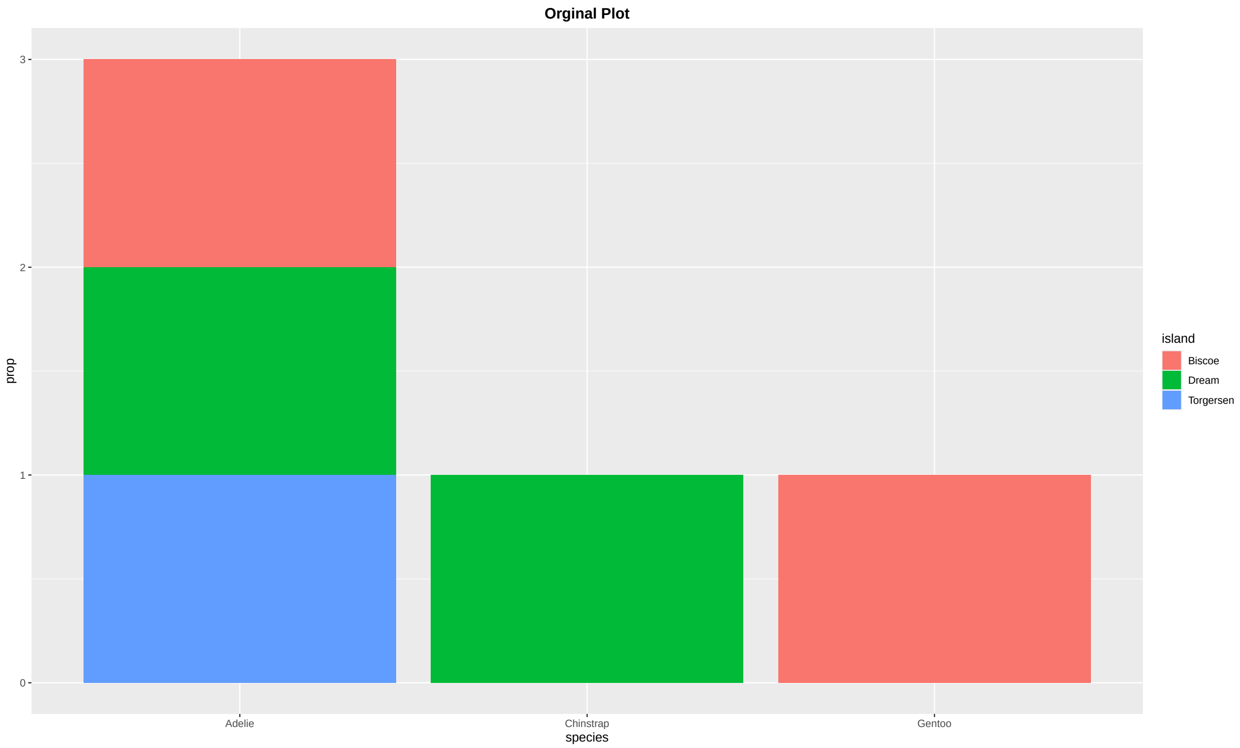
Task 3: Comparison of geom_point() and geom_count()

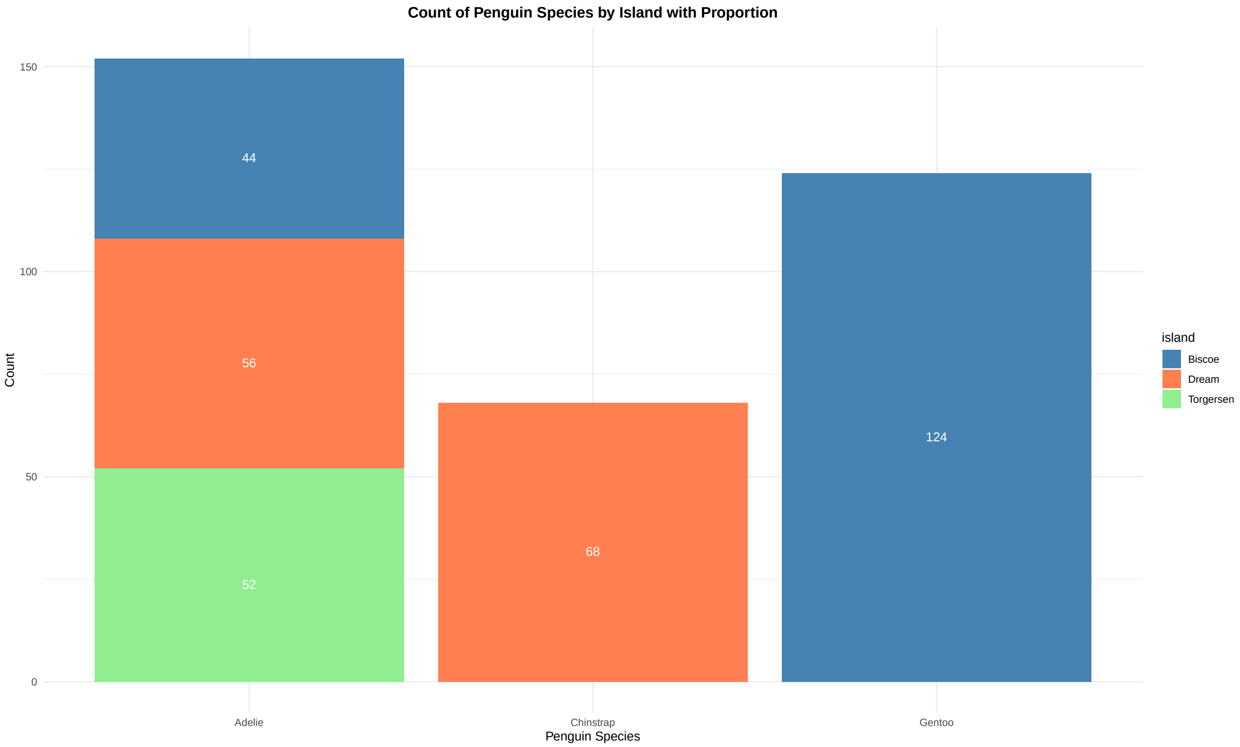


Explanation

For Task 3, we use geom_point() and geom_count() to explore the relationship between cty (city MPG) and hwy (highway MPG) in the mpg dataset. In the scatter plot with geom_point() (as suggested in the slides on visualizing distributions),each point represents an observation, which can obscure dense areas if points overlap. Adding slight transparency (alpha) mitigates this and reveals overlapping data points. In contrast, geom_count() changes the size of each point based on its count,providing a clear indication of where data points are densest, a recommendation seen in the slides for visual clarity when values overlap. Overall, geom_point() is ideal for datasets with low overlap, while geom_count() effectively highlights density in datasets with repeated values. These enhancements meet the general assignment criteria for readability and clarity in data presentation.

Task 4: Penguins

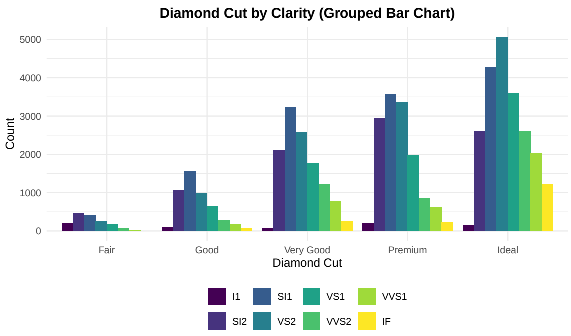


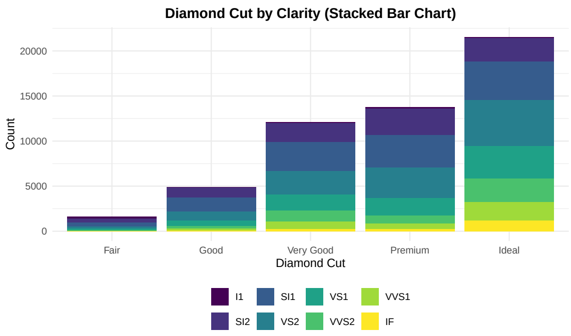


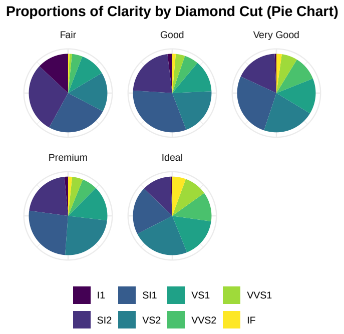
Problems with the original plot

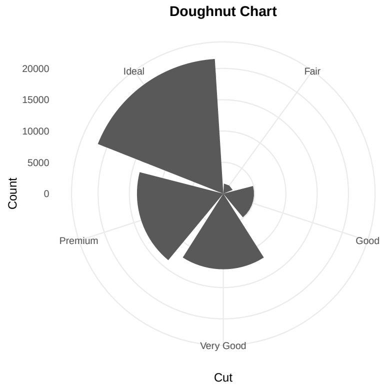
In the original plot, you can only see which penguin species are present on each island. For example, on Torgersen Island, there are only Adelie penguins, whereas Biscoe Island has both Gentoo and Adelie penguins. However, it does not give a sense of where there are larger populations or the total number of birds on each island. That's why we added the total counts to each bar in our updated plot. Additionally, the proportions are now displayed accurately, which enhances the clarity of the visualization. For instance, you can immediately see that while Adelie penguins are the most common species, Biscoe Island has the largest overall penguin population.

Task 5: Diamond









Comparison

For this task, we created four visualizations to explore the distribution of clarity within each cut category in the diamonds dataset. The grouped bar chart displays clarity levels side-by-side for each cut category, allowing straightforward comparisons of diamond counts by clarity, as seen on slide 10. The stacked bar chart also shows the same data but stacks clarity levels within each cut category. This visualization highlights the relative contributions of each clarity level to the total counts, enhancing our understanding of the dataset's composition. Next, the faceted pie chart provides insights into clarity proportions for each cut, displaying one pie chart per cut. This format illustrates the distribution of clarity levels clearly, utilizing polar coordinates as advised on slides 12 and 13. Finally, the alternative pie chart view transforms a basic bar chart into a pie chart using polar coordinates. While less detailed than the faceted version, it offers a quick overview of clarity proportions across all cut categories. In summary, the grouped bar chart excels for clear comparisons, while the stacked bar chart and both pie charts offer valuable insights into the relative proportions of clarity, each showcasing its unique strengths.

Workload

We sat down at the beginning and divided up the tasks. The first thing we did was set up a Git repository so that we could easily collaborate and continuously track each other's progress. We organized the tasks so that Fabian took on tasks 1 and 2, while Samuel handled tasks 3-5. This clear division helped us focus and avoid overlapping efforts. After completing our initial assignments, we held a short meeting to discuss the status and share updates on our progress. This check-in proved beneficial, as it highlighted some areas that needed further refinement. Although a few items were still incomplete, we felt that we were moving in the right direction and understood what remained to be done. To ensure the quality of each other's work, we decided that Fabian would review and make corrections to Samuel's tasks, and vice versa. This mutual review process not only helped catch errors but also facilitated a better understanding of each other's approach. By the end of the session, we felt more confident about our progress and looked forward to wrapping up the remaining tasks.