ECO395M: Data Mining — Exercise 01

Scott Cohn

2021-02-07

Data visualization: Gas Prices

```
# boxplot
gasprices %>%
  ggplot( aes(x=Competitors, y=Price, fill=Competitors)) +
    geom_boxplot() +
    scale_fill_brewer(type="qual") +
    geom_jitter(color="black", size=0.6, alpha=0.9) +
    theme_minimal() +
    theme(text=element_text(family="Palatino")) +
    labs(y="Price")
```

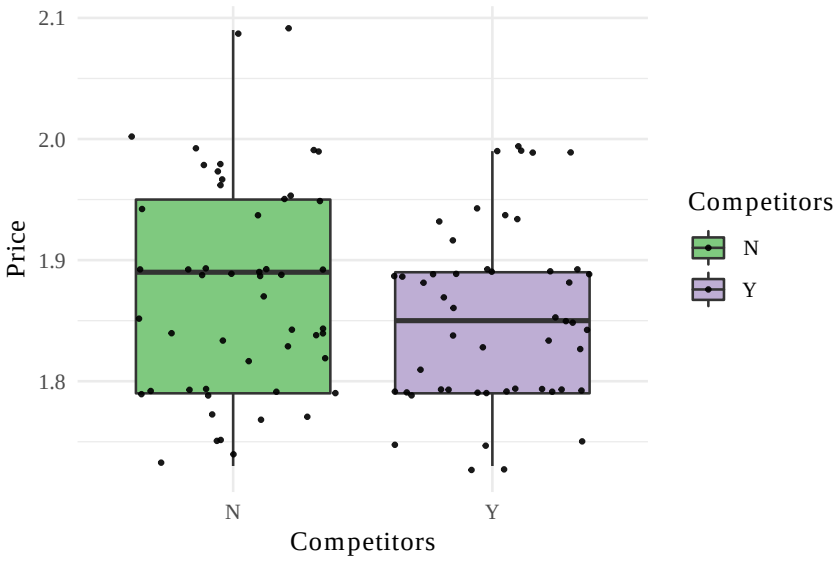

Figure 1: Price with competition

THEORY A: Gas stations charge more if they lack direct competition in sight.

CONCLUSION: The cheapest gas stations seem to charge similarly. Gas stations that have no competition have higher prices at the 50th and 75th percentile. There are several gas stations, again facing no competition, pricing 10 cents per gallon higher than all of the rest.

```
# scatter plot
gasprices %>%
  ggplot(aes(x=Income, y=Price, color=Brand)) +
  geom_jitter() +
  scale_x_continuous(breaks=seq(10000,130000,15000)) +
  scale_color_brewer(palette="Set1") +
  geom_rangeframe(color="black") +
  theme_tufte()
```

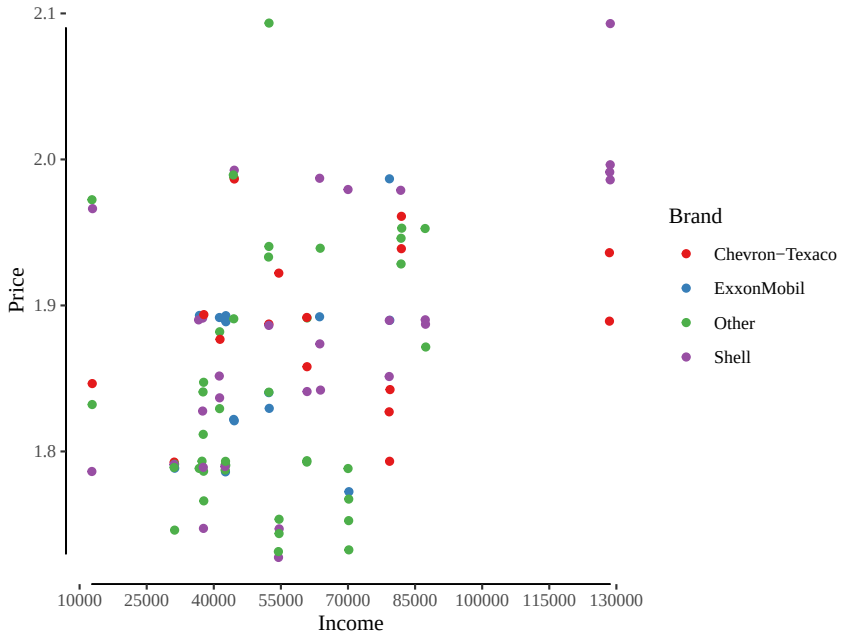

Figure 2: Price with income

THEORY B: The richer the area, the higher the gas price.

CONCLUSION: There is a lot of variation within brand across the income spectrum. There does seem to be a trend that higher income areas do have higher prices per gallon at the gas station.

```
# Bar plot
gasprices %>%
  ggplot( aes(x=Brand, y=Price, fill=Brand) ) +
  stat_summary(fun.data=mean_sdl, geom="bar") +
  stat_summary(fun.data=mean_cl_boot, geom="errorbar", width=0.3) +
  labs(x="") +
  coord_flip() +
  scale_fill_brewer(palette="Set1") +
  theme_tufte()
```

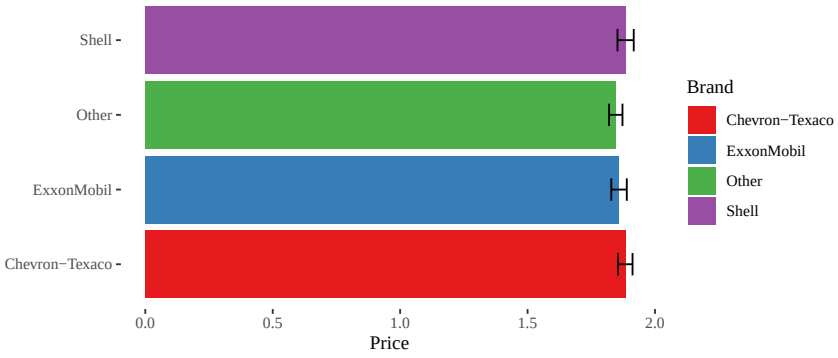

Figure 3: Shell versus competition

THEORY C: Shell charges more than other brands.

CONCLUSION: Visually, the averages are very close; Shell seems to price similarly to other brands. I would not conclude, visually, that Shell prices notably higher than the rest.

```
# faceted histogram
gasprices %>%
  ggplot( aes(x=Price, fill=Stoplight) ) +
  geom_histogram(position="identity", bins=15) +
  facet_grid(. ~ Stoplight) +
  scale_fill_brewer(type="qual") +
  labs(y="Count") +
  geom_rangeframe() +
  theme_tufte()
```

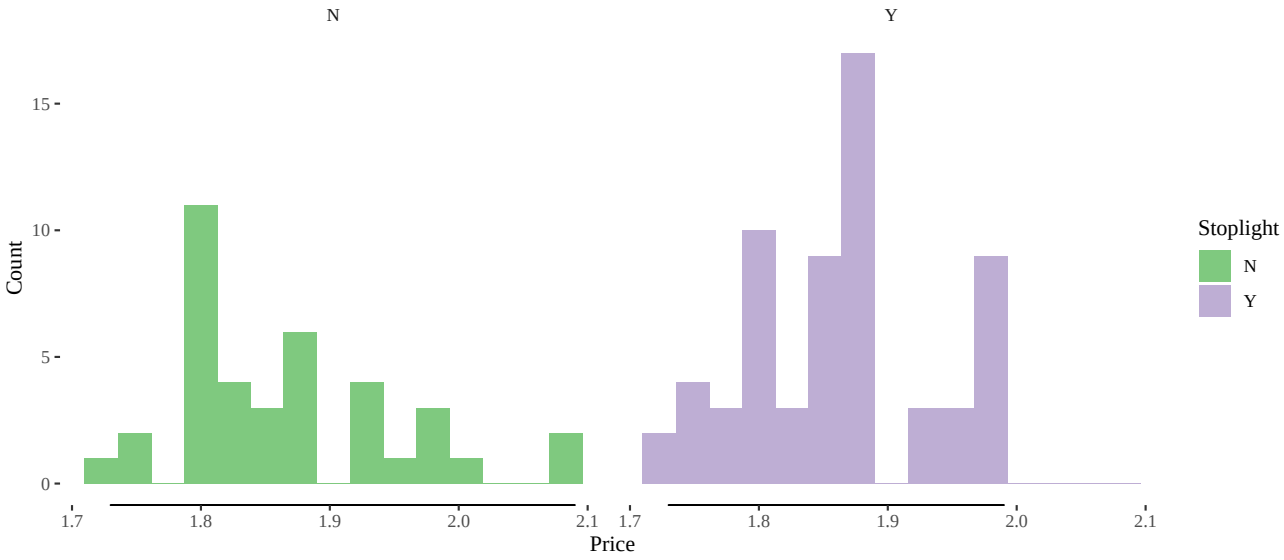

Figure 4: Price with stoplights

THEORY D: Gas stations at stoplights charge more.

CONCLUSION: The distribution of price not at a stoplight is right-skewed whereas the distribution of price at a stoplight appears to be more normally distributed (maybe even left-skewed). Limited data makes this a bit difficult to parse. They seem to be similarly centered around 1.8ish. It seems that gas stations as stoplights price a bit higher than their non-stoplight counterparts.

```
# any
p1 <- gasprices %>%
  ggplot( aes(x=Price, fill=Highway)) +
    geom_density(alpha=0.6) +
    scale_fill_brewer(type="qual") +
    theme_tufte() +
    labs(y="Density")

p2 <- gasprices %>%
  ggplot( aes(x=Highway, y=Price, fill=Highway)) +
    geom_boxplot() +
    scale_fill_brewer(type="qual") +
    geom_jitter(color="black", size=0.4, alpha=0.9) +
    theme_tufte() +
    theme(legend.position="none") +
    labs(y="Price")

# patchwork
(p2 | p1)
```

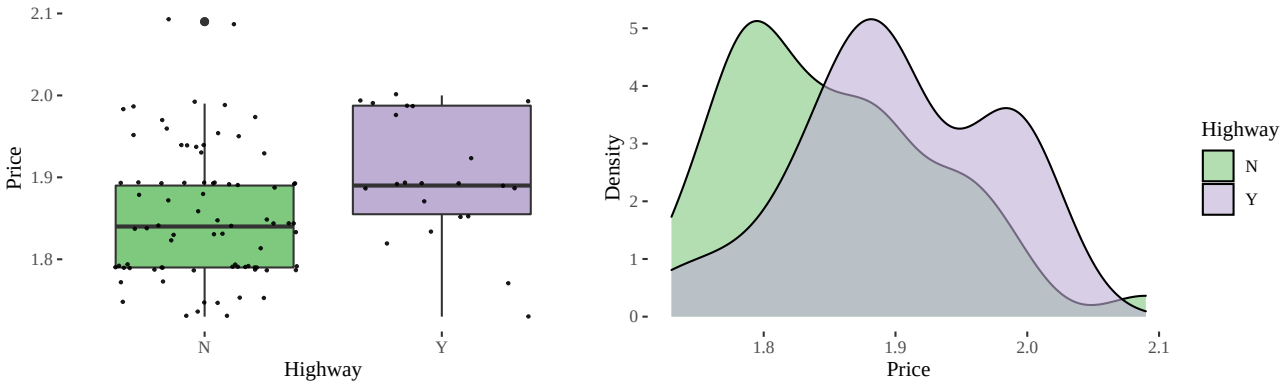

Figure 5: Price with highway access

THEORY E: Gas stations with direct highway access charge more.

CONCLUSION: The distribution of prices without direct highway access is more left-skewed than the price distribution with direct access. This is clear to see in the density plot on the right, but also when comparing the medians in the boxplot on the left.

Also, we can make a map by geocoding the addresses.

```
tryCatch({
  # Try: Run map
  atx_map <- ggmap(get_map("austin", zoom = 13),
                   ylab="Latitude",
                   xlab="Longitude")

  m2 <- atx_map +
    geom_point(data=gasprices,
               mapping = aes(x=longitude,
                             y=latitude,
                            color=Brand)) +
    scale_color_brewer(palette="Set1")

  m2 + labs(x="Longitude",
            y="Latitude") +
    theme(text = element_text(family="Palatino"))
  # Catch: If replicating, need unique google API to work
}, error = function(e) {
  "Error: Need Google map API to run map code for ggmap"
  }
)
```

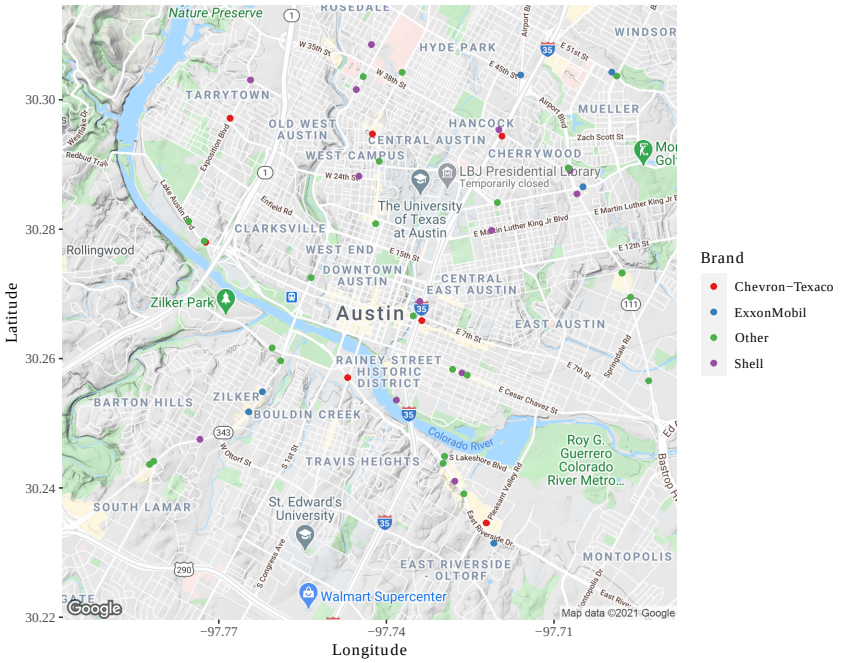

Figure 6: Map of Austin gas stations

Data Visualization: A Bike Share Network

Plot 1

```
p1 <- bikeshare %>%
  group_by(hr) %>%
  mutate(avg_tot = mean(total)) %>%
  ggplot( aes(x = hr) ) +
  geom_jitter(aes(y = total), color = palette_light()["blue"], alpha = 0.15)

p2 <- p1 + geom_line( aes(y = avg_tot), color = palette_light()["red"])

p3 <- p2 +
  labs(x = "Hour of day (24hr)",
       y = "Average rentals") +
  theme_tufte()

p3
```

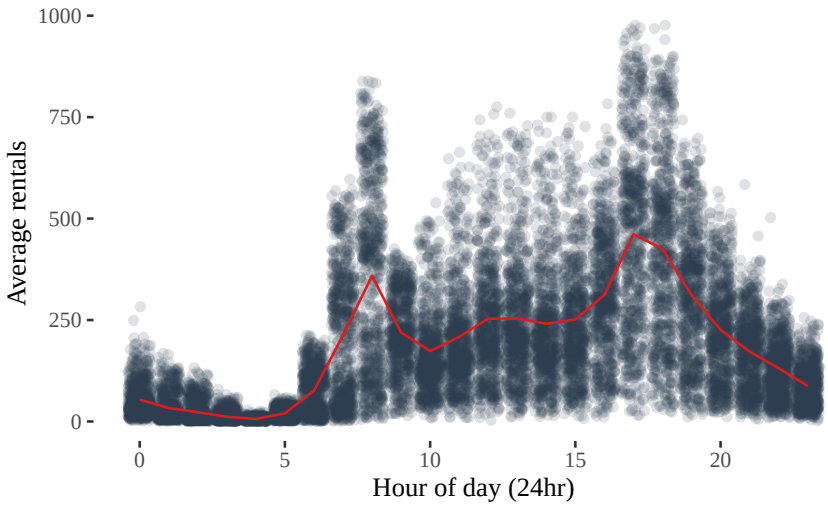

Figure 7: Bike rentals (total) versus hour of day (hr)

This plot shows the total bike rentals per day by hour (24hr). The red line shows the averages by hour, and is smooth to connect across discrete hours counts. The points are "jittered" to show density. It seems bike rentals peak before and after work hours. This might indicate folks using the bikes to commute to and from work.

Plot 2

```
workingday_labs <- c("Not working day", "Working day")
names(workingday_labs) <- c(0, 1)

p2 <- p1 + geom_smooth(aes(y = total), color = palette_light()["red"]) +
  facet_grid(. ~ workingday,
                labeller = labeller(workingday = workingday_labs))

p2 +
  labs(x = "Hour of day (24hr)",
       y = "Average rentals") +
  theme_tufte()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

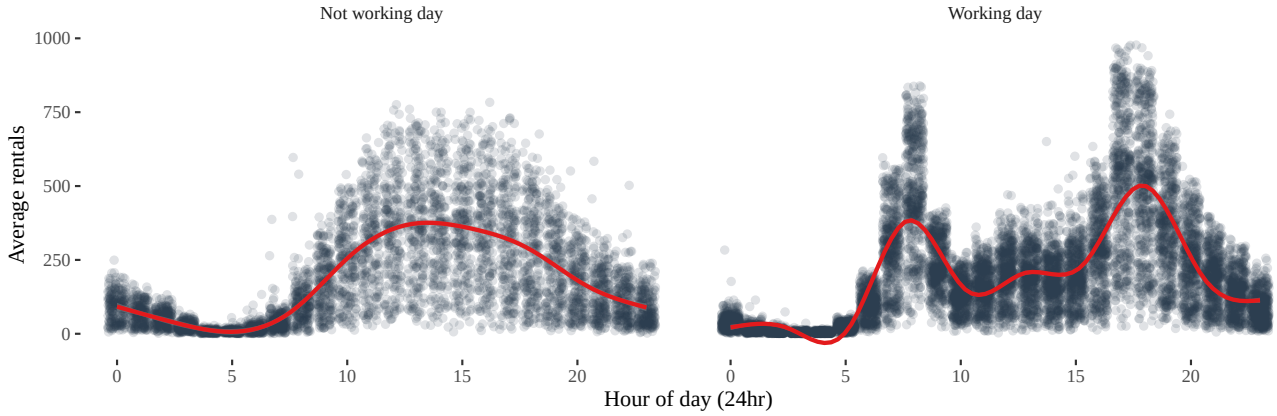

Figure 8: Bike rentals (total) versus hour of day (hr) by working day

The design of this plot is identical to the previous, except now it is faceted by whether it is a working day or not. On non-working days, we see a "normal distribution" of bike rentals peaking in the early afternoon. On working days, we see the bimodal humps just before work and just after; again, this likely indicates bike rentals for commuting to and from work.

Plot 3

```
g1 <- bikeshare %>%
    filter(hr == 8) %>%
    group_by(weathersit) %>%
    mutate(avg_tot = mean(total)) %>%
    ggplot() +
    geom_col(aes(x = weathersit, y = avg_tot,
                 fill = factor(weathersit)))

weathersit_labs <- c("Clear", "Mist", "Light Snow")

g2 <- g1 +
  scale_y_continuous(breaks = seq(0, 160000, by = 25000)) +
  labs(x = "Weather Situation", y = "Average rentals") +
  theme_tufte() +
  scale_fill_brewer(palette = "Set1", labels = weathersit_labs) +
  theme(legend.title = element_blank(),
        legend.position = "bottom",
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank())

g2 + facet_grid(. ~ workingday,
        labeller = labeller(workingday = workingday_labs))
```

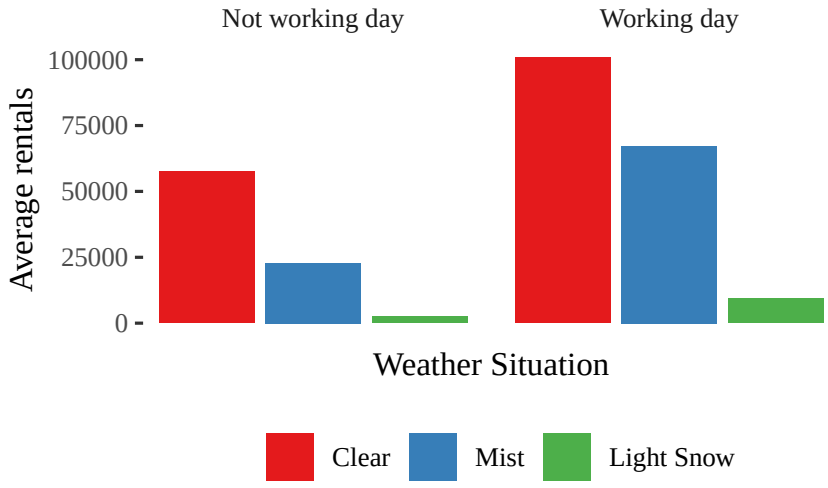

Figure 9: Average ridership during 8 AM hour by weather code and working day

This plot breaks down bike rentals by weather. As expected, improvements in weather are associated with increased bike rentals, with less total rentals on non-working days.

Data visualization: Flights at ABIA

Here, we visualize the flightpaths available in the dataset. We only observe domestic flights. Each vertex represents an airport, and each edge is a flight between Austin Bergstrom (green) and another airport.

In an effort to narrow down the dataset, we want to find which carriers complete 90% of the flights between Austin and any of these other destinations. We find that 10 out of 16 carriers complete 90% of flights. For these purposes, 90% is largely arbitrary, a more careful analysis would consider the impact of dropping these 6 carriers. For example, maybe a dropped carrier provides all of the flights to a particular region. We do not address those concerns here.

Table 1: Top carriers by number of flights

UniqueCarrier	n	perc_flights	cum_sum
WN	34876	0.3513601	0.3513601
AA	19995	0.2014407	0.5528007
CO	9230	0.0929881	0.6457888
YV	4994	0.0503123	0.6961011
B6	4798	0.0483377	0.7444388
XE	4618	0.0465243	0.7909631
OO	4015	0.0404493	0.8314125
OH	2986	0.0300826	0.8614951
MQ	2663	0.0268285	0.8883236
9E	2549	0.0256800	0.9140036

A carrier delay is where the cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.). What is the probability of carrier delay by an airline?

Here, we show the probability of carrier delay by day of week. Mondays and Fridays have the highest likelihood of a carrier delay.

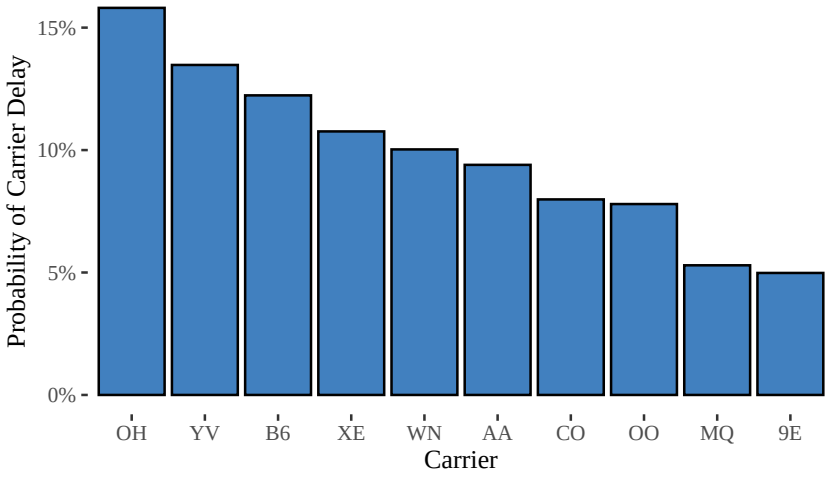

Figure 10: Probability of delay by carrier

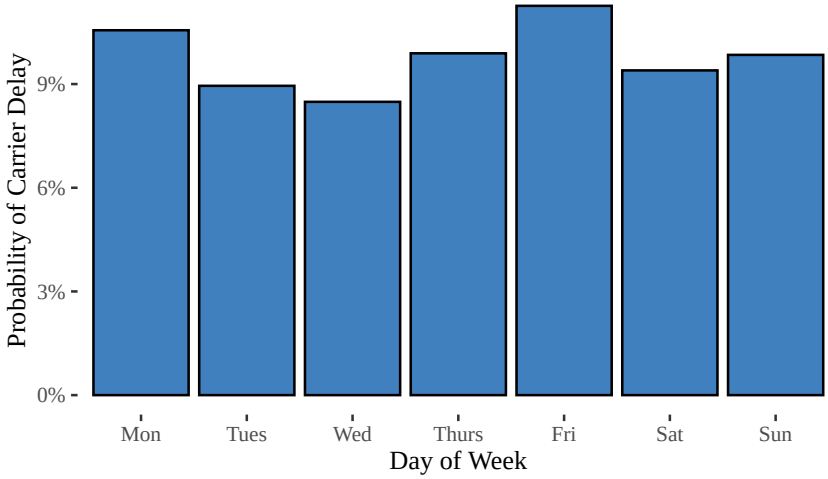

Figure 11: Probability of carrier delay by day of week

It seems that flying on Mondays and Fridays is associated with the highest probability of encountering a carrier delay. We could change the `delay` variable and repeat this analysis for all other delays if we wish.

K-Nearest Neighbors

Table 2: Count by trim

trim	n
350	416
65 AMG	292

sclass 350

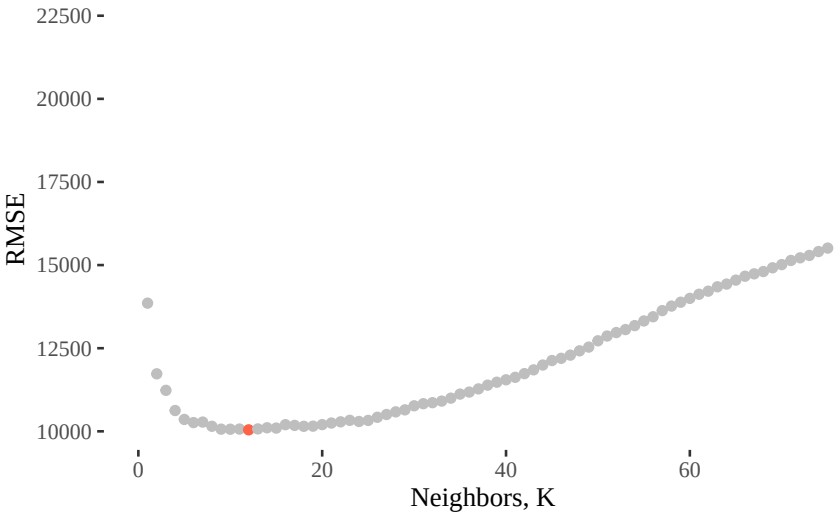

Figure 12: RMSE versus k

Table 3: Minimum RMSE

neighbors	.metric	.estimator	mean	n	std_err	.config
12	rmse	standard	10044.45	2	214.4052	Preprocessor1_Model012

```
## [1] "The smallest rmse occurs at K = 12"
```

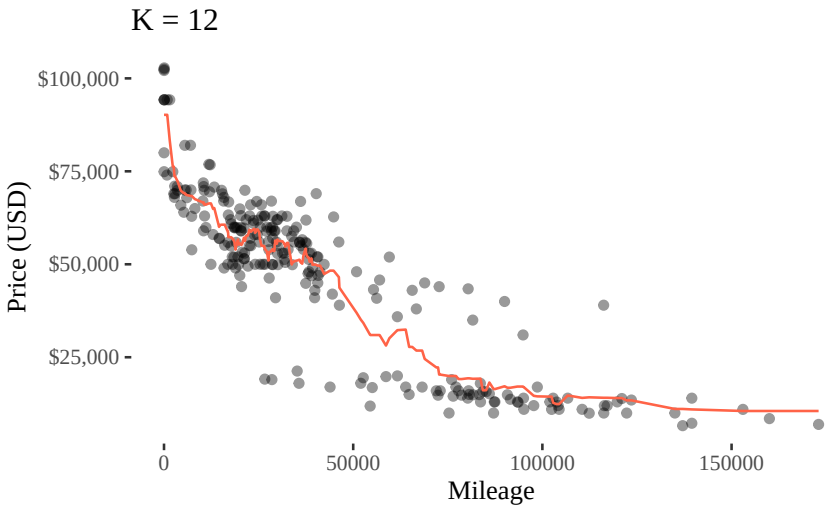

Figure 13: Overlaid fitted model for optimal value of K, where red is the 65 AMG trim and blue is the 350 trim.

sclass 65 AMG

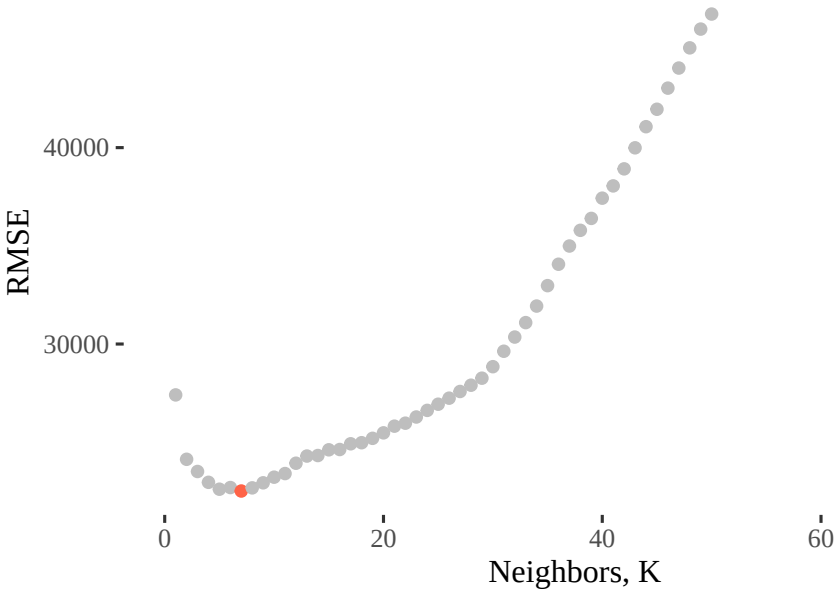

Figure 14: RMSE versus K

Table 4: Minimum RMSE

neighbors	.metric	.estimator	mean	n	std_err	.config
7	rmse	standard	22514.13	2	1853.518	Preprocessor1_Model07

```
## [1] "The smallest rmse occurs at K = 7"
```

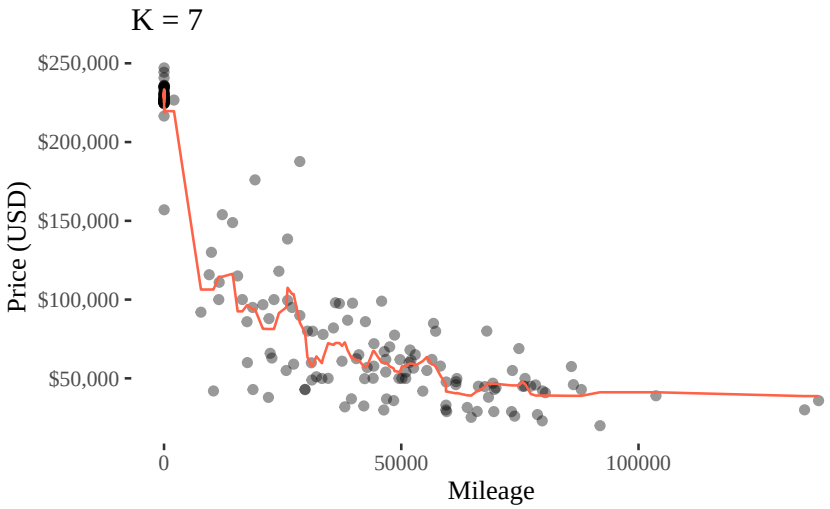

Figure 15: Fitted model for optimal value of K

The sclass 350 needs a higher value of K because more points might lead to overfitting for a lower value of K.

It is also interesting to look at both `mileage` versus `price` graphs together.

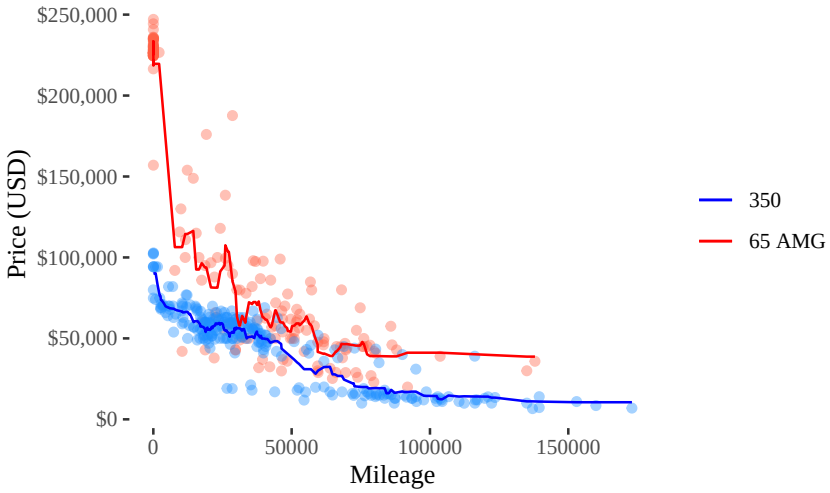

Figure 16: Fitted model for optimal value of K