Group P: Project Analysis Component

Sydney Bluestein, Parker Smith, Tate Eppinger, and Juan Perez

2024-04-03

```
Cars <- janitor::clean_names(Sport_car_price)
Cars <- unique(Cars)
Cars <- Cars %>%
  filter(!grepl("Electric", engine_size_l)) %>%
  mutate(across(engine_size_l:x0_60_mph_time_seconds, as.numeric))
Cars2<-Cars[-c(94, 150, 154, 409),] #deletes middle eastern cars and super old car
Cars3<-filter(Cars2, price_in_usd<500000) #removed data from IQR test below
Cars4<-na.omit(Cars3)
## MAR: Recategorized into "Asian", "Domestic", and "European"
SportsCars <- Cars4 %>% mutate(car_make = case_when(
    car_make %in% c("Acura", "Kia"  , "Lexus"  , "Mazda"  , "Nissan" , "Subaru" , "Toyota") ~ "Asian",
    car_make %in% c("Ford" , "Chevrolet" , "Dodge") ~ "Domestic",
    car_make %in% c("Alfa Romeo" , "Alpine" , "Ariel" , "Aston Martin" , "Audi" , "Bentley", "BMW", "Bu
```

Does “what’s under the hood” matter when predicting sports car prices in America? vroom vroom Group P: Sydney Bluestein, Tate Eppinger, Juan Perez, and Parker Smith

Project Aim: We want to explore which factors contribute to sports car prices in America. We have chosen numerous explanatory variables that may influence sports car prices: car origin, car year, 0-60 speed, torque, horsepower, and engine size.

Research Question: Do any of the factors above (or a combination of them) predict sports car prices accurately?

Variables: Car origin (Asian, Domestic, or European) is our categorical variable. Year is a quantitative variable that represents the year the car was made. 0-60 speed is a quantitative variable that represents the time in seconds it takes for the car to go from 0 miles per hour to 60 miles per hour. Torque is a quantitative variable measured in pounds-feet that describes the power of the car’s drivetrain. Horsepower is another quantitative variable measured in foot-pounds/minute representing engine power. Lastly, engine size is a quantitative variable in liters that measures how much gas the car can hold.

Step 1 – Univariate Analysis: We included a brief univariate analysis in our project proposal. We included density plots of horsepower, torque, engine size, and 0-60. It would be useful to obtain descriptive stats for these quantitative variables using the favstats method. Looking into the quantitative variables’ shape, center, and spread will help orient us toward recognizing typical and unusual values for variables measured in unfamiliar units. Car year is characterized by a box plot. We removed a few cars produced before 2014 since we are attempting to predict sports car prices from the last decade. We also included a bar graph for our categorical variable, car origin. We notice that European cars are most commonly represented in this data set. However, there are still enough Asian and domestic sports cars to make this variable worthwhile. We will use the tally function to create a table illustrating the exact distribution of cars originating from Europe, Domestic, and Asia.

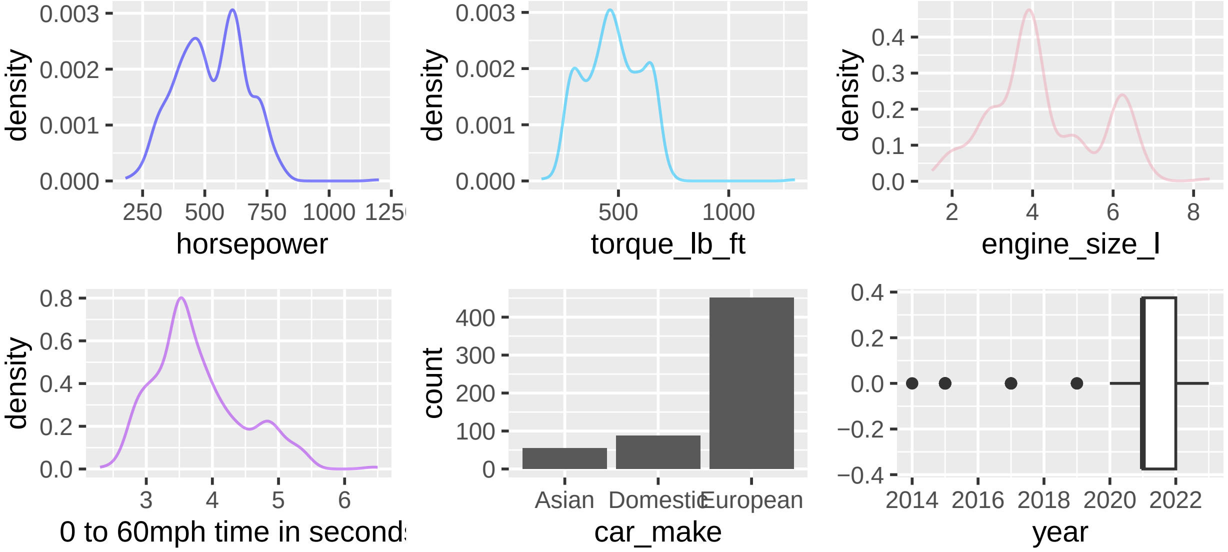
```
m1<-gf_dens(~ horsepower, data = SportsCars, color = "blue")
m2<-gf_dens(~ torque_lb_ft, data = SportsCars, color = "deepskyblue")
m3<-gf_dens(~ engine_size_l, data = SportsCars, color = "pink2")
```

```
m4<-gf_dens(~ x0_60_mph_time_seconds, data = SportsCars,  xlab="0 to 60mph time in seconds", color = "pu
```

```
m5<-gf_bar (~car_make, data= SportsCars)
m6<-gf_boxplot(~year, data=SportsCars)
favstats(~year, data=SportsCars) #IQR test 1.5, 1.5(IQR=1)=1.5, data range from 2019 to 2023
```

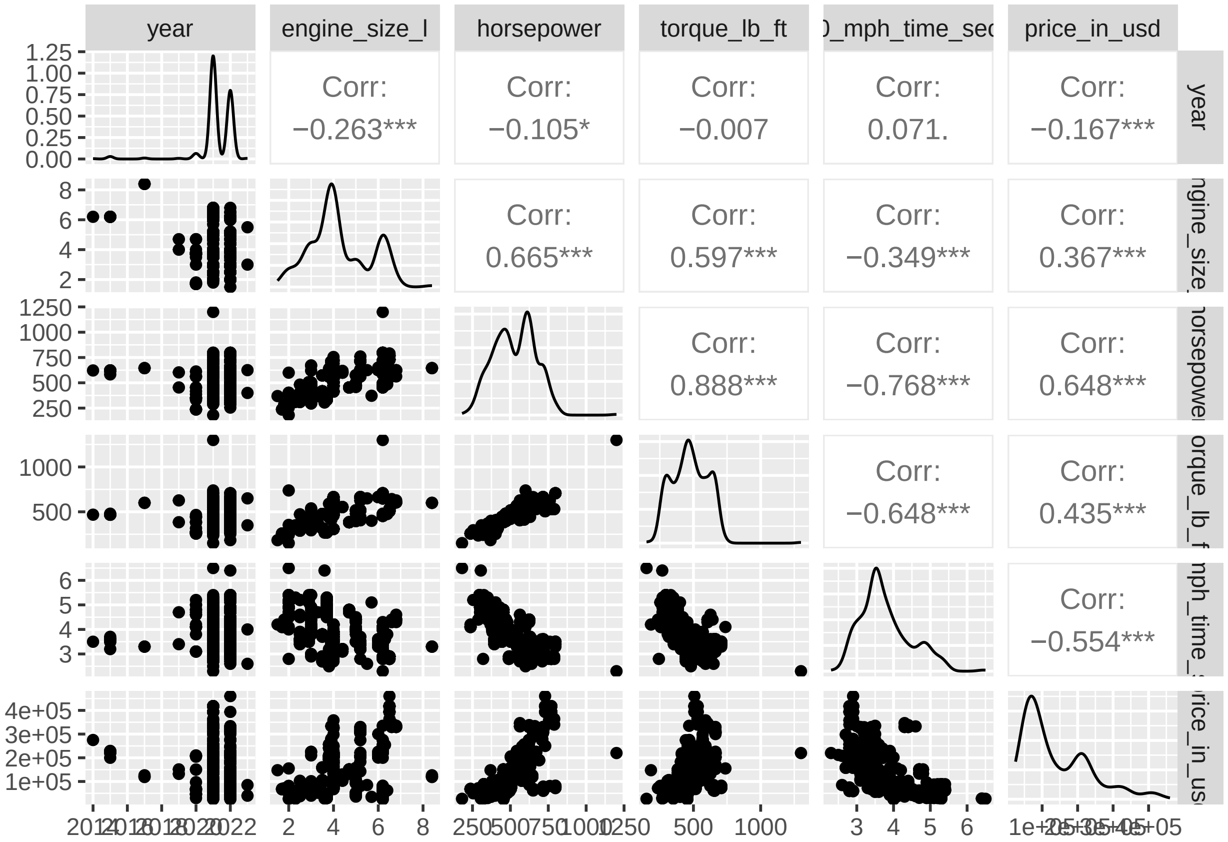
```
##   min   Q1 median   Q3  max     mean       sd   n missing
##  2014 2021   2021 2022 2023 2021.233 1.015807 592       0
```

```
grid.arrange(m1, m2, m3, m4, m5, m6, ncol=3, heights=c(2, 2, 2))
```



Step 2 – Bivariate Analysis: We will begin with simple linear regression to gauge which explanatory variables help predict sports car prices. It will be time-efficient to begin with GGpairs as a summary of the relationships between our explanatory variables and sports car prices. Upon the first glimpse, the GGpairs plot demonstrates that horsepower (r=0.648) and 0-60 (r=-0.554) may be helpful predictors for sports car prices. We plan to explore this relationship further by creating simple linear regression models. To check for linearity, we can look either at the SLR plot or the residuals vs. fitted plot to ensure the relationship between our predictor variable and sports car price is best described as linear. Independence is not a condition we can assume from a plot. (We have very little information regarding the origin of this dataset, so we will have to proceed with caution). Looking at whether the data points fit well on the expected line QQ plot line will allow us to decide whether the normality condition is satisfied. We will confirm homoscedasticity for the equal variance condition by looking for a seemingly random distribution of points on the residuals vs. fitted plot. If we encounter any issues with conditions, we will look into removing potential outliers or transforming the data if necessary.

```
ggpairs(SportsCars, columns = c(3:8))
```



Step 3 – Model Building and Variable Selection: Once we have looked into SLR, we will turn to multiple linear regression. Looking at the GGpairs plot, we notice many potential pairs of predictors run the risk of being collinear. (torque and horsepower, 0-60 and horsepower, and horsepower and engine size). We can run a VIF test if we choose to include variables that may be collinear to confirm we have created a stable model. We may use an automated variable selection method (maybe stepwise, so we get the best predictors at every stage) to delineate which combination of explanatory variables will help predict sports car prices best. We will then weigh the models' parsimony with adjusted R-squared values and Mallow's Cp to determine the best model(s). If we have a potential outlier, we can use a Cook's distance plot to determine if the point is unusual enough to remove from our model.

Regarding our categorical variable, we would like to run an ANOVA test to determine if the mean sports car price differs based on its region of origin. Suppose the ANOVA test is associated with a significant F statistic. We will then run a Tukey test (or a similar multiple comparison test) to determine which region(s) of origin have significantly different sports car prices. The results from this multiple comparison test will help us determine how we need to adjust our model, i.e., for what region of origin do we need to account for higher sports car prices relative to the other regions? Perhaps we would want to look into some interaction terms as well.