Project Proposal, Group P 02/25/24 Title: Group Members: Group P–Sydney Bluestein, Parker Smith, Tate Eppinger, and Juan Perez

Title: Does "what's under the hood" matter when predicting sports car prices in America? *vroom vroom*

Research Question/Purpose: Who doesn't love a cool sports car? Parker's dad and Sydney's grandfather are particularly interested in car shows/collecting. Therefore, we grew up listening to our fair share of "why this car will be worth millions one day." We've always wondered, what makes these cars so much more expensive than the cars we drive daily? Does the make and model of the car matter, or is the price based on the practicality and performance of the vehicle? Is a Ferrari more expensive because of the name brand or because it can go from 0-60 faster than other car brands? In general, we would like to investigate the relationship between predictors such as car make and horsepower on race car price. We are interested in assessing our hypothesis that name brands contribute more to race car prices than practical measures such as torque. We also want to investigate the relationship between year and car price. We expect that car price would be high for "antique cars," but new cars would also have high prices. This would lead to a curvature in the data. So, would we have to transform the data to make it linear? Obviously, these are very early predictions/ideas, and I look forward to determining if the evidence supports or contrasts our initial inclinations.

Data: We pulled data from the Kaggle Data Explorer, updated a year ago. (It appears like Kaggle is not very highly regulated, so we should proceed with caution). This data set was created to identify market trends relating to sports car prices. The author of this dataset did not mention how they selected which cars to include in the set; therefore, we should proceed with caution. We have reached out to "KIATTISAK RATTANAPORN" the author of this data set for more information.

Population: This data set contains sports cars from various manufacturers worldwide. Therefore, it would not be wise to expand this model to all cars or car manufacturers outside of those included in the data set. The IQR includes cars produced in the years 2014-2019. Thus, we should not predict the price of cars produced anywhere outside of this range. The data set from Kaggle is a sample from sports cars in America, so we hope to be able to use this model as a way to predict the price of sports cars in the US so that hopeful owners can get the right car for the right price.

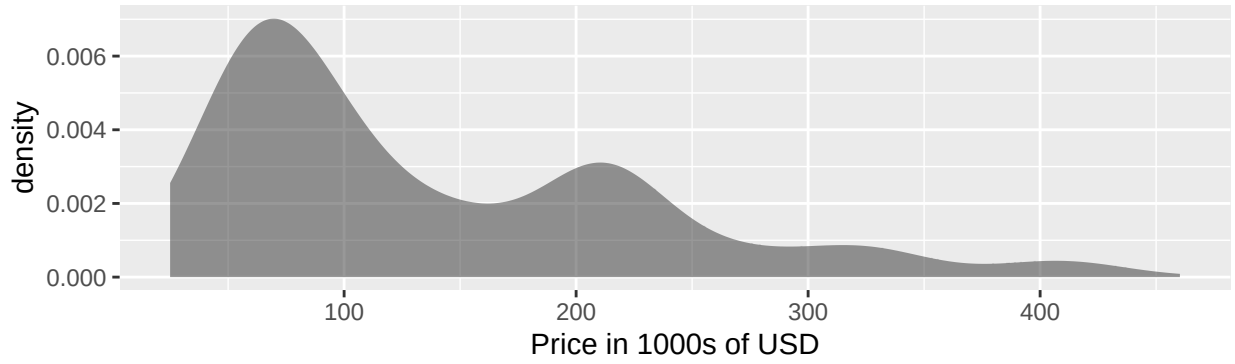
Response Variable: Our response variable is the price of sports cars. This is a quantitative variable that is measured in US dollars. Sports car price is not normally distributed, as the data is skewed right. The median sports car price is $102,550. The IQR is ~ 140,000 dollars.

Explanatory Variables: We have numerous explanatory variables to look into from the Kaggle data set. First, car origin (Asian, Domestic, or European) is our categorical variable. Origin is represented in a bar graph. 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 measures the power of the car's drive train. Horsepower is another quantitative variable measured in foot-pounds/minute that represents the engine power. Lastly, engine size is a quantitative variable in liters that measures how much gas the car can hold. We included distributions of our explanatory variables below.

```
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", "Bug
```

```
gf_density(~price_in_usd/1000, data=SportsCars, xlab="Price in 1000s of USD") #specify 1000s, response
```




```
favstats(~price_in_usd, data=SportsCars) #IQR test 1.5, 1.5(IQR=152720)=229080, data range from $298980
```

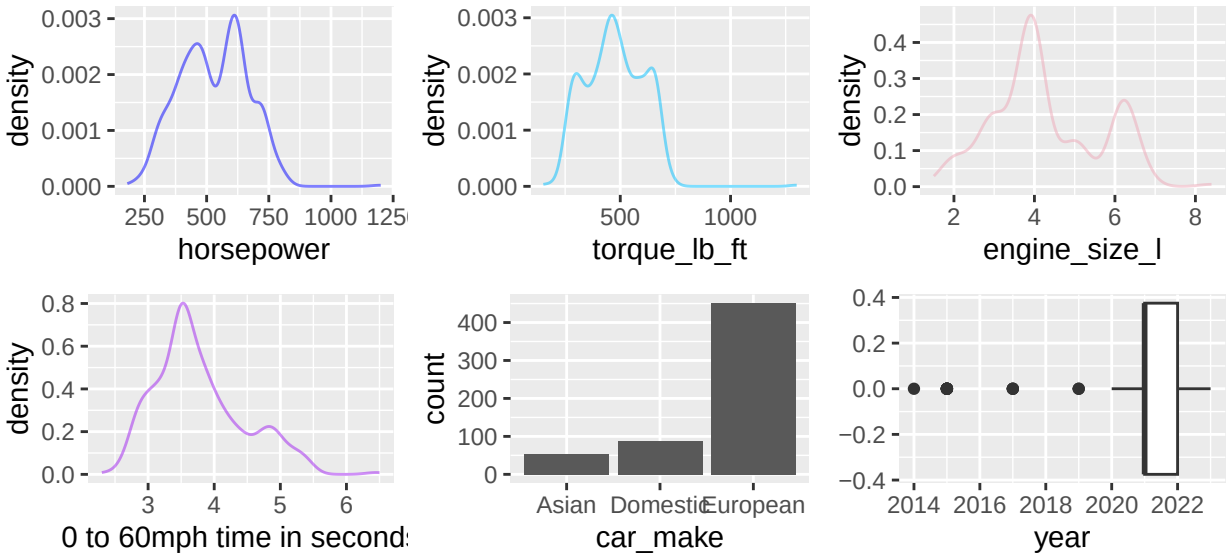
```
##    min    Q1 median       Q3    max     mean       sd   n missing
##  25000 67575 102550 207206.2 460247 139567.3 93941.05 592       0
```

```
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))
```


density
0.003
0.002
0.001
0.000
250
500
750
1000
horsepower
torque_lb_ft
engine_size_l
0.4
0.3
0.2
0.1
0.0
2
4
6
8
0.8
0.6
3
5
0 to 60mph time in second
count
400
300
200
100
0
Asian
Domestic
European
car_make
−0.2
−0.4
2014
2016
2018
2020
2022
year