DATA 606 Data Project Proposal

Vic Chan

Data Preparation

```
# load data
dataset = read.csv('https://www.fueleconomy.gov/feg/epadata/vehicles.csv')

# Filtering out all options that are unnecessary

library(tidyr)
library(dplyr)

dataset = dataset %>%
  filter(fuelType == 'Regular' |  fuelType == 'Premium' | fuelType == 'Midgrade') %>%
  select(city08, highway08, cylinders, displ, drive, fuelType, make, model, year, trany)
```

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

Have fuel economy increased throughout the years for gas engine cars or have they peaked? This means that we will be needing to filter out diesel, electric, and hybrid vehicles as they will impact the study.

Cases

What are the cases, and how many are there?

Each row represents a vehicle make and model. In the dataset there are 43,418 different vehicle make and models spanning from 1984 to 2021.

Data collection

Describe the method of data collection.

The data is collected from The US Department of Energy and includes all make and models of vehicles from 1984 to 2021.

Type of study

What type of study is this (observational/experiment)?

This is a observational study

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

U.S. Department of Energy, Energy Information Administration, Independent Statistics & Analysis. (2021, March 25). Fuel Economy Data. Retrieved from https://www.fueleconomy.gov/feg/download.shtml

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

The response variable is miles per gallon city/highway and it is quantitative.

Independent Variable

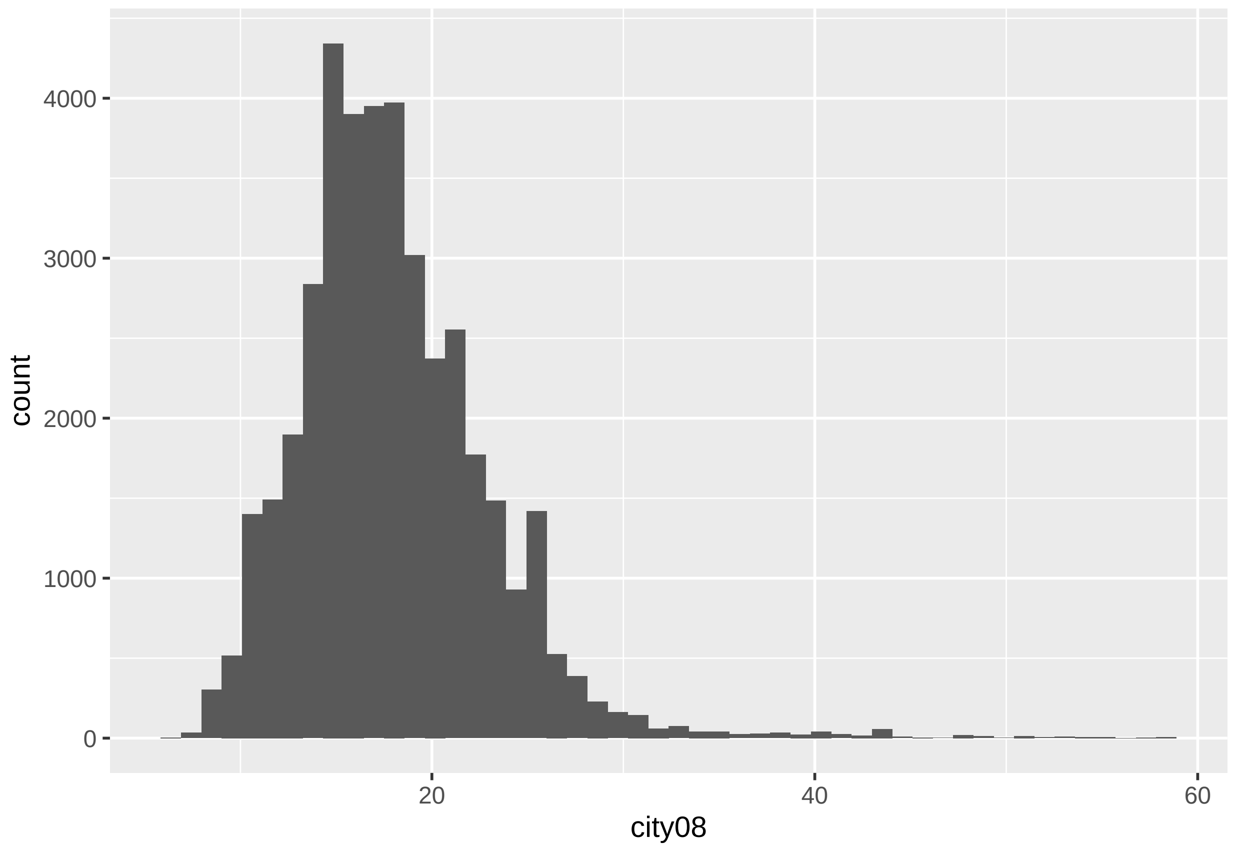
You should have two independent variables, one quantitative and one qualitative.

Quantitative - Cylinders Qualitative - Drive Train Qualitative - Transmission Type Quantitative - Displacement Quantitative - Year

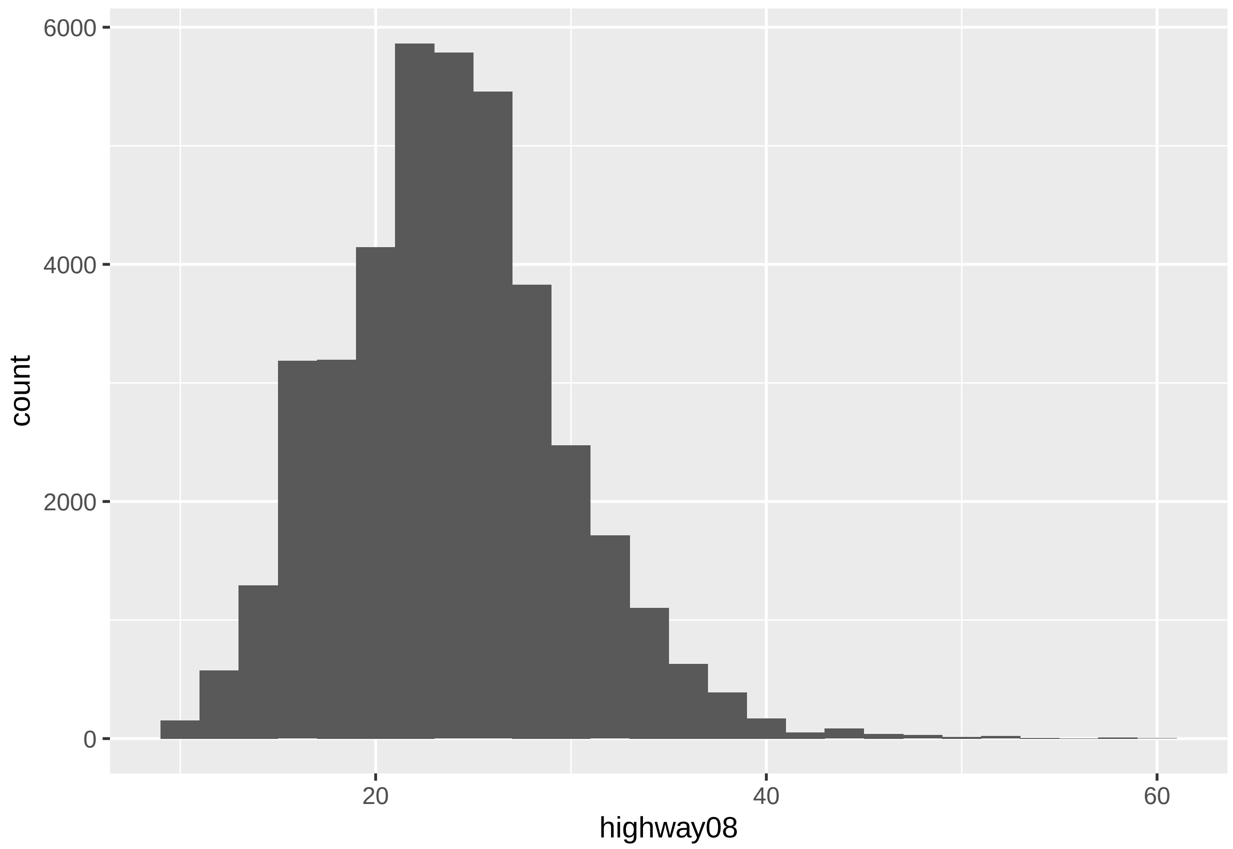
Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

```r
library(ggplot2)

dataset %>%
  ggplot(aes(x=city08)) + geom_histogram(bins = 50)
```



```
dataset %>%
  ggplot(aes(x=highway08)) + geom_histogram(bins = 27)
```



```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(dataset[,c("city08", "highway08", "displ", "cylinders")])
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 2 rows containing missing values
```

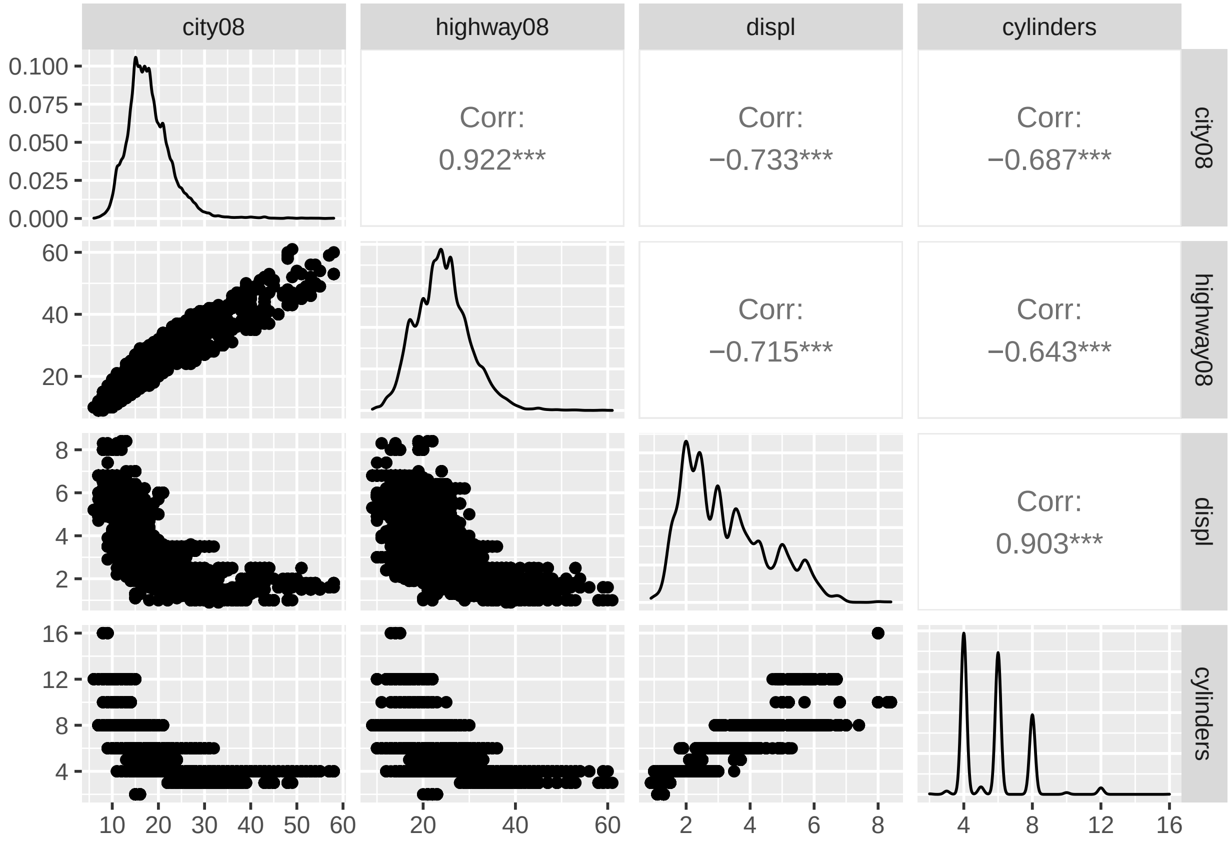
```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 3 rows containing missing values
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 2 rows containing missing values
```

```
## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 3 rows containing missing values
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
## Warning: Removed 2 rows containing non-finite values (stat_density).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 3 rows containing missing values

## Warning: Removed 3 rows containing missing values (geom_point).

## Warning: Removed 3 rows containing missing values (geom_point).

## Warning: Removed 3 rows containing missing values (geom_point).

## Warning: Removed 3 rows containing non-finite values (stat_density).
```


city08
highway08
displ
cylinders
Corr:
0.922***
Corr:
−0.733***
Corr:
−0.687***
Corr:
−0.715***
Corr:
−0.643***
Corr:
0.903***
city08
highway08
displ
cylinders