Homework Assignment 1

Vasyl Ostapenko (774 970 8)

April 02, 2022

1. Supervised learning uses datasets with labeled classes and we can judge model accuracy in performing classification or regression. Unsupervised learning uses datasets without class labels and thus we ask a model to find clusters (patterns) on its own without knowing true accuracy.
2. A model used for a regression task will predict a continuous quantity and thus will output a real value for a sample. A model used for a classification task will predict a discrete class label and thus will output a binary (or discrete for multiclass problems) value after a cutoff is used.
3. In regression, we might commonly use MSE or RMSE as a measure of model performance. In classification, we might instead use accuracy or F1-score to measure model performance.
4. A descriptive model summarizes the data without interpreting it. Inferential models use hypothesis testing and confidence intervals to know more about population parameters of interest when one has some sample from that population. Finally, a predictive model learns relationships in existing data and uses them to make a prediction on a new sample or samples.

5a. Empirical models describe data with very few assumptions about the data (not assuming much about f). Mechanistic models describe data after specifying assumptions (such as parametric form for f) and trying to incorporate known factors about the systems surrounding the data into the model. A mechanistic model requires some knowledge of the system's structure and function and needs many features. On the other hand, an empirical (statistical) model is used when the system is extremely complex and has many unknowns. Thus we seek a mathematical description of the system based on external characteristics.

5b. A mechanistic model is easier to understand because we understand the underlying system and can interpret the parameters and their relationships and roles.

5c. A mechanistic model might suffer from high bias because we make many assumptions about the data (system). An empirical (statistical) model might suffer from high variance because we seek many features to describe the relationships in the data (system).

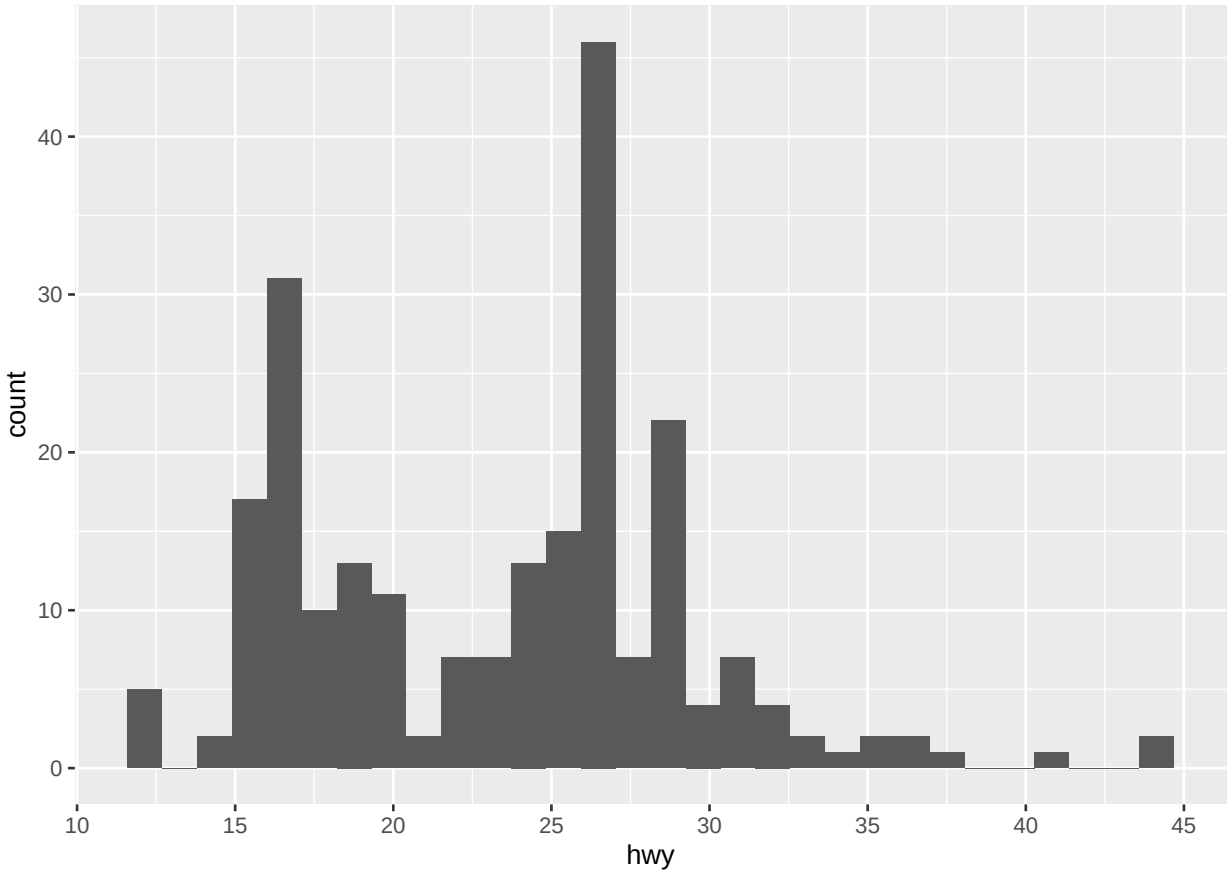
6. The first question, formulated using the words "how likely", is inferential. This is because assessing predictive quality is an inferential act. On the other hand, the second question, formulated using the wording "how would", is purely predictive. This is because we seek to predict the future given some predictors.

EDA

```
data = mpg
```

Ex1.

```
ggplot(data, aes(x=hwy)) +
  geom_histogram(bins=30) +
  scale_x_continuous(breaks=seq(0, 50, 5))
```

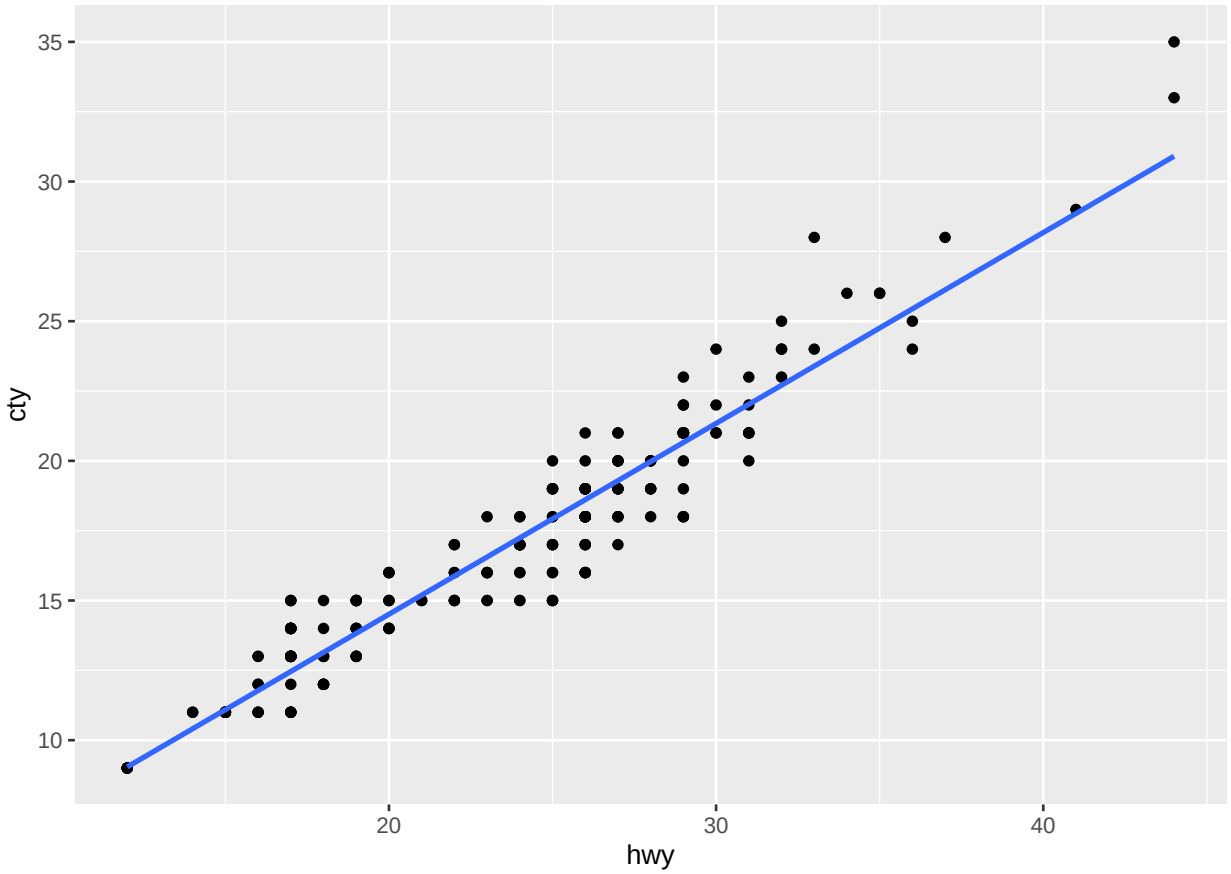


We see a bimodal distribution for the hwy variable, with peaks around 16 and 26 mpg. The first peak may account for trucks and SUVs, while the second peak may account for sedans and coupes.

Ex2.

```
ggplot(data, aes(x=hwy, y=cty)) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE)
```

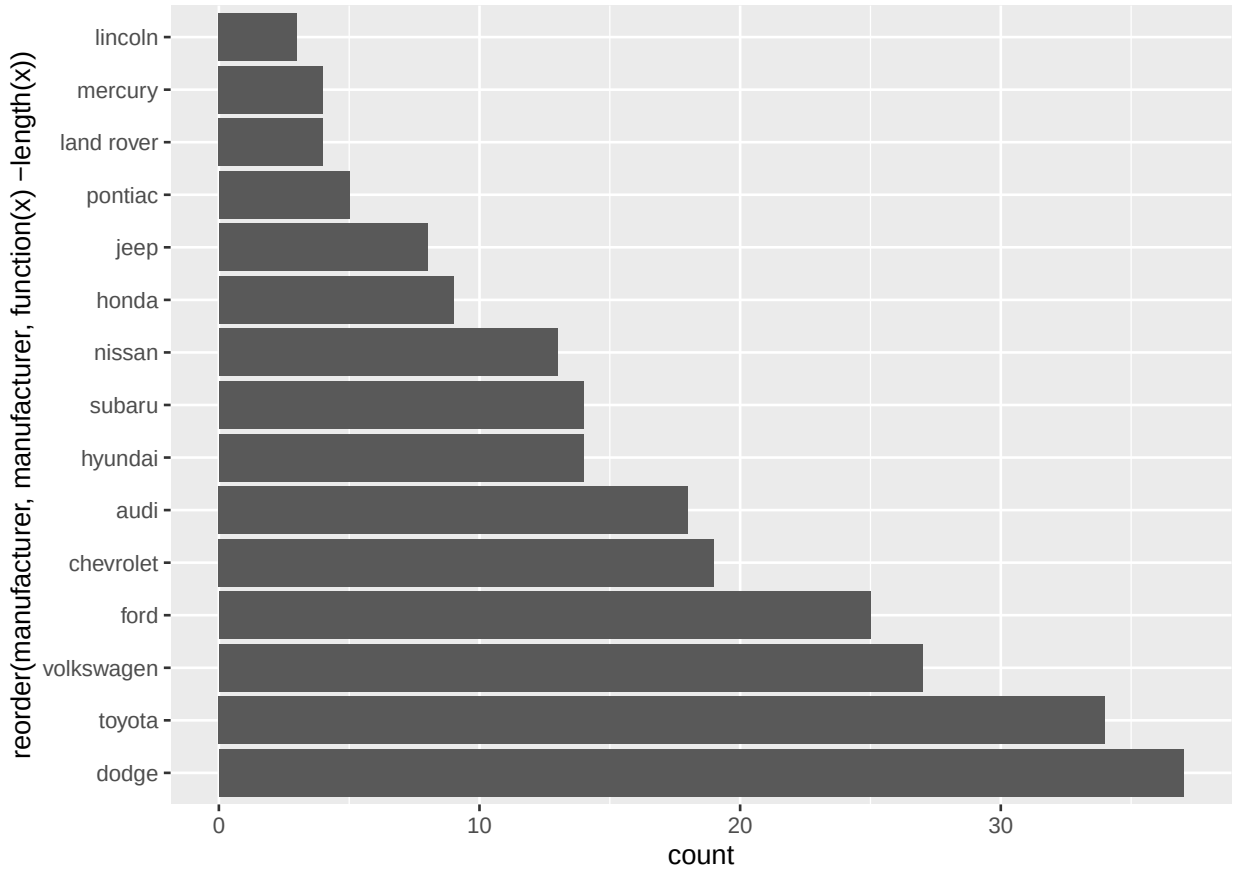
```
## `geom_smooth()` using formula 'y ~ x'
```



It looks like there is a strong linear relationship between highway and city mpg values. Logically, this makes sense, because a car with high fuel economy will maintain this both on city roads and on highways.

Ex3.

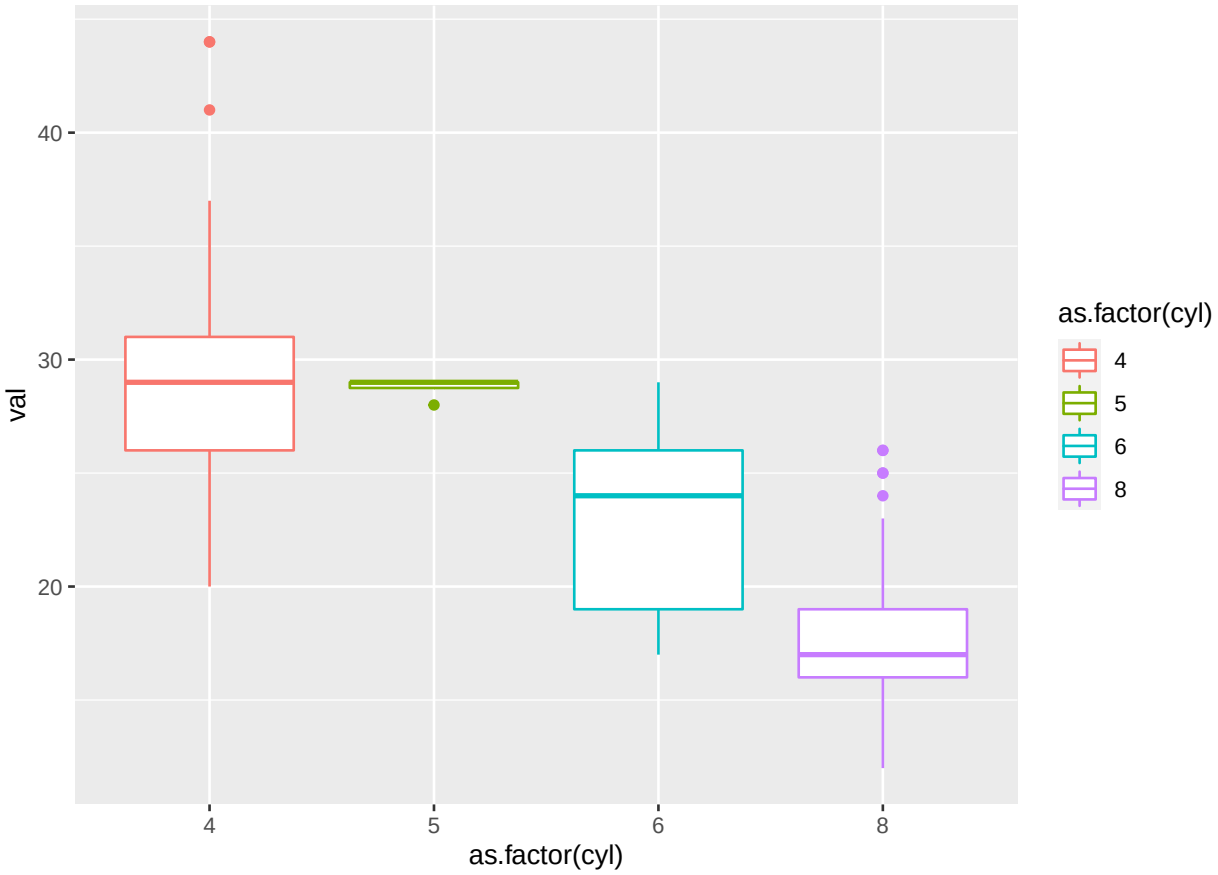
```
ggplot(data, aes(x=reorder(manufacturer, manufacturer, function(x)-length(x)))) +
  geom_bar() +
  coord_flip()
```



Lincoln produced the fewest cars, while Dodge produced the most.

Ex4.

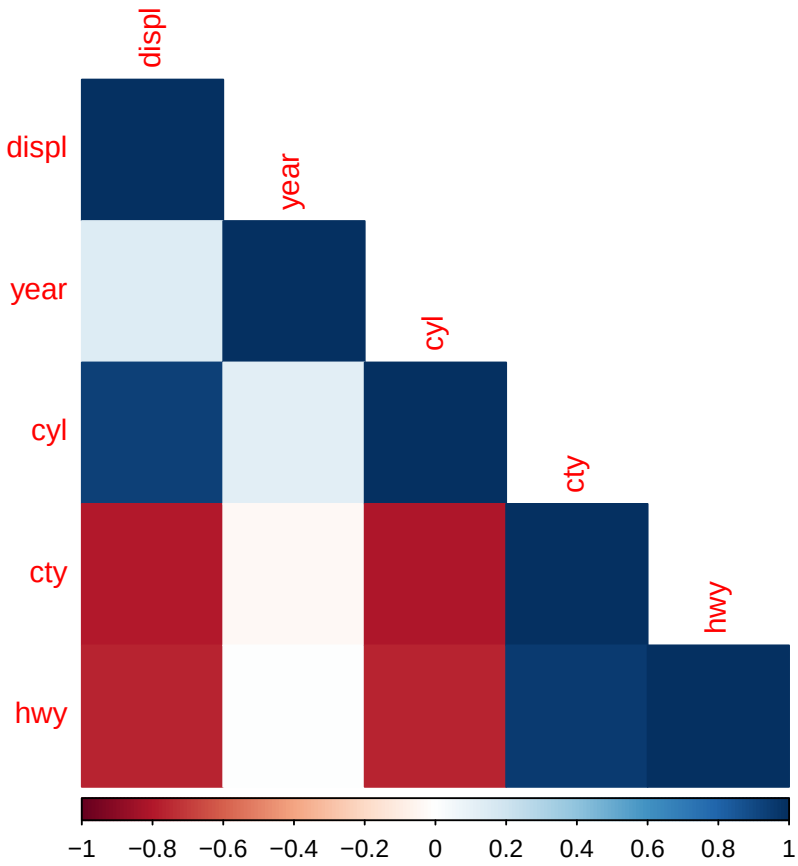
```
data2 = as.data.table(data)
data2 = melt(data2, measure.vars="hwy", value.name="val", variable.name="var")
ggplot(data2, aes(x=as.factor(cyl), y=val, color=as.factor(cyl))) +
  geom_boxplot()
```



Higher numbers of cylinders tend to be associated with lower values for highway mpg.

Ex5.

```
data3 = data[ , sapply(data, is.numeric)]
corrplot(cor(data3), method="color", type="lower")
```



City and highway mpg are both negatively correlated with engine displacement and number of cylinders. This makes sense, because larger, more powerful engines tend to be less fuel efficient. City and highway mpg are strongly correlated. This makes sense as we stated in exercise two. Finally, number of cylinders and engine displacement are strongly correlated. This makes sense because larger engines need more cylinders to drive the car.