Assignment 1 Solution

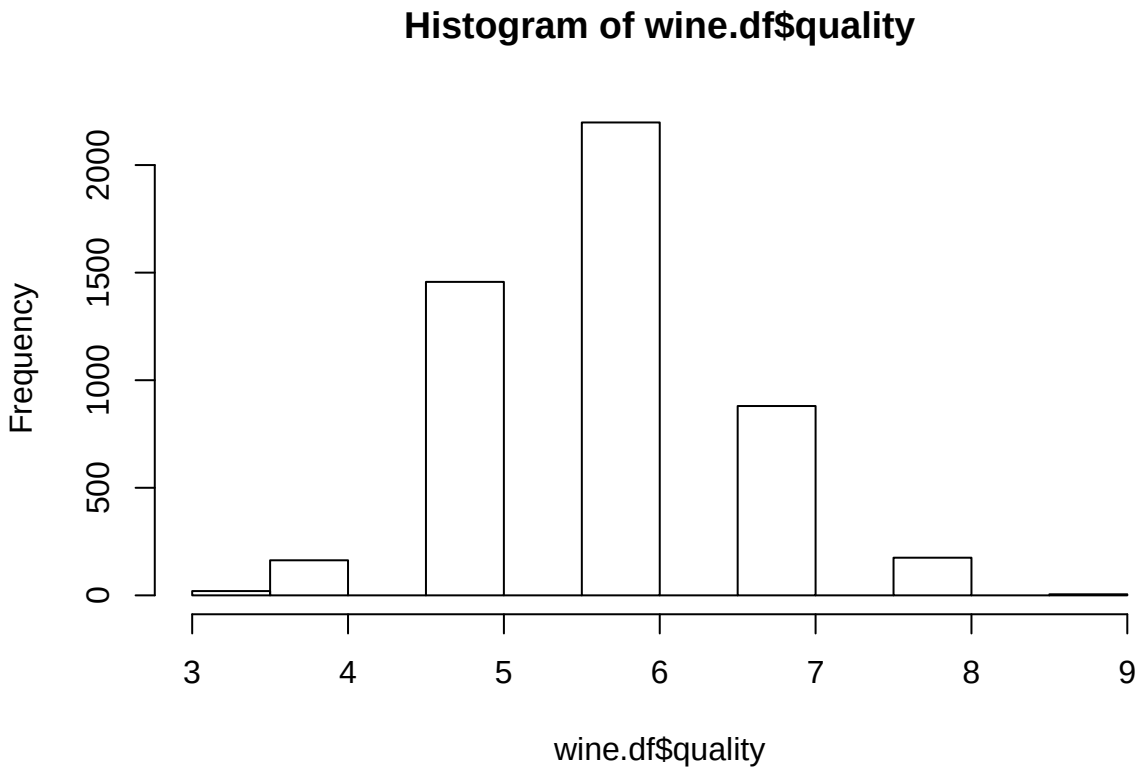
4/20/2018

Preperation

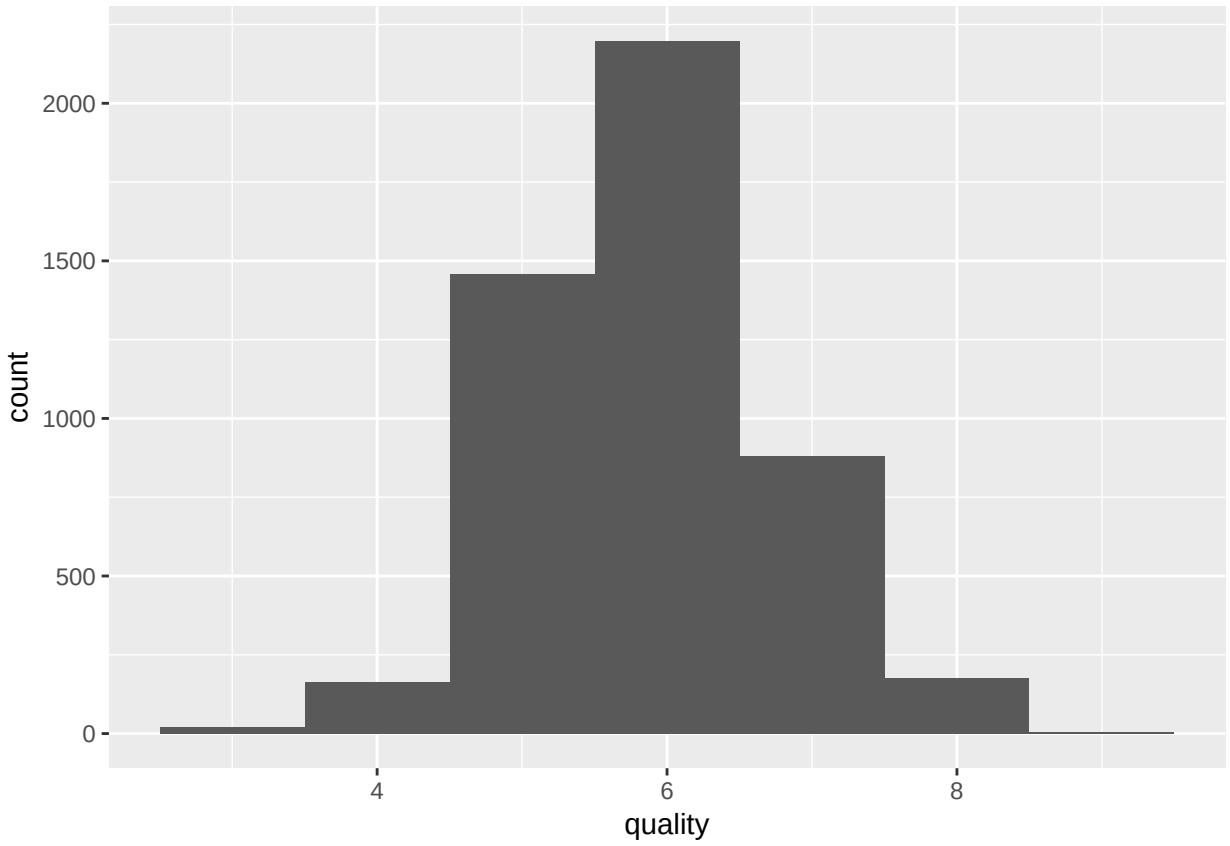
```
# read data
wine.df <- read.csv("wine.csv")
```

1. Plot a histogram for the outcome variable.

```
hist(wine.df$quality)
```




```
##alternate graph using ggplot
library(ggplot2)
ggplot(data = wine.df) +
  geom_histogram(aes(x = quality),binwidth = 1)
```



2. Partition the data into 70% Training and 30% Validation with a seed of 555

```
set.seed(555)  # set seed for reproducing the partition
# use nrow to get the total number of observations
nobs <- nrow(wine.df)
train.index <- sample(1:nobs, 0.7 * nobs)
train.df <- wine.df[train.index, ]
valid.df <- wine.df[-train.index, ]
```

3. Use the training set to run a multiple linear regression for "quality" vs. all other predictors.

```
wine.lm <- lm(quality ~ ., data = train.df)
```

4.1 Which predictors are statistically significant?

```
options(scipen = 999)
summary(wine.lm)
```

```
##
## Call:
## lm(formula = quality ~ ., data = train.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5742 -0.4933 -0.0372  0.4616  3.1647
```

```
##
## Coefficients:
##                          Estimate   Std. Error t value
## (Intercept)           122.3135393   21.1454329   5.784
## fixed.acidity           0.0258613    0.0244060   1.060
## volatile.acidity       -1.7773902    0.1349200 -13.174
## citric.acid             0.0280990    0.1160554   0.242
## residual.sugar          0.0709722    0.0086445   8.210
## chlorides              -0.4223139    0.6590883  -0.641
## free.sulfur.dioxide     0.0028851    0.0009923   2.908
## total.sulfur.dioxide   -0.0004114    0.0004534  -0.908
## density              -121.6990218   21.4661132  -5.669
## pH                      0.5513568    0.1247470   4.420
## sulphates               0.6050679    0.1211574   4.994
## alcohol                 0.2190064    0.0272810   8.028
##                                 Pr(>|t|)
## (Intercept)           0.00000000793263418 ***
## fixed.acidity                     0.28939
## volatile.acidity     < 0.0000000000000002 ***
## citric.acid                       0.80870
## residual.sugar        0.00000000000000031 ***
## chlorides                         0.52173
## free.sulfur.dioxide               0.00367 **
## total.sulfur.dioxide              0.36420
## density               0.00000001552365363 ***
## pH                    0.00001018635363642 ***
## sulphates             0.00000062083828290 ***
## alcohol               0.00000000000000135 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7516 on 3416 degrees of freedom
## Multiple R-squared:  0.2719, Adjusted R-squared:  0.2695
## F-statistic:   116 on 11 and 3416 DF,  p-value: < 0.00000000000000022
```

- As we can see, the following predictors are statistically significant:
  - volatile.acidity
  - residual.sugar
  - density
  - pH
  - sulphates
  - alcohol

4.2 How do you interpret the effect of "density" based on the estimates?

The estimates for density is -121.6990218, which means an increase of 1 unit in density will cause a decrease of -121.6990218 of in the dependent variable, quality.

5. What is the RMSE(or residual standard error) for the training set?

We can use `wine.lm$fitted.values` to retrieve the fitted values of the training set.

```
library(forecast)
```

```
## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone 'zone/tz/2018c.
```

```
## 1.0/zoneinfo/America/Los_Angeles'
```

```
accuracy(wine.lm$fitted.values, train.df$quality)
```

```
##                                      ME      RMSE       MAE       MPE     MAPE
## Test set -0.000000000000000000564054 0.7502895 0.5833139 -1.734718 10.29482
```

The RMSE is 0.741614.

6.1 Apply the model to the validation set and assess the performance. What is the RMSE for the validation set?

```
wine.lm.pred <- predict(wine.lm, valid.df)
accuracy(wine.lm.pred, valid.df$quality)
```

```
##                    ME      RMSE       MAE       MPE     MAPE
## Test set 0.02741676 0.7527734 0.5874028 -1.311781 10.30643
```

The RMSE is 0.772009 for validation set.

6.2 Do you think there is overfitting problem?

No. If our model does *much better* on the training set than on the validation set, then we're likely overfitting. But in our model, the RMSE for training and validation set are in similar scale. So it is unlikely that we are overfitting.