The Impact of Outliers: The Raisin River Canoe Race | SOLUTIONS

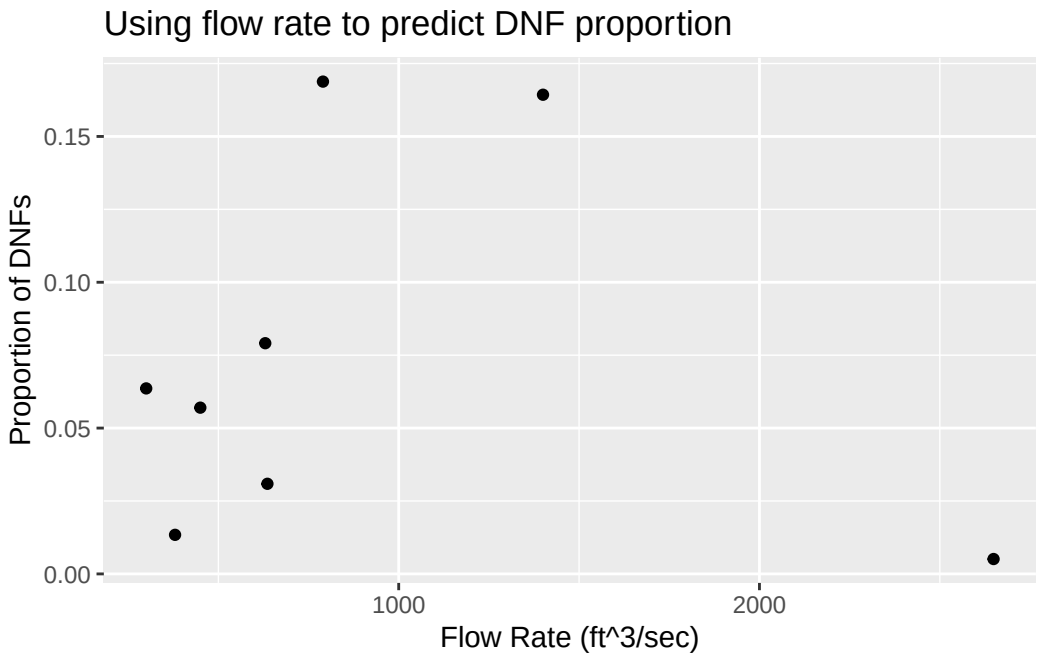
```
library(tidyverse)
library(dplyr)
library(broom)

RR_dnf_data <- read_csv("raisin_river_DNF.csv")
```

Explore the relationship between flow rate and proportion of DNFs

1. Graph a scatterplot that shows flow rate as a predictor for proportion of DNF

```
ggplot(RR_dnf_data, aes(x = flow, y = prop_DNF)) +
  geom_point() +
  labs(x = "Flow Rate (ft^3/sec)",
       y = "Proportion of DNFs",
       title = "Using flow rate to predict DNF proportion")
```



2. Fit a model using flow rate as a predictor for proportion of DNFs, print a summary of the model, and meaningfully interpret the slope coefficient.

```
dnf_mod <- lm(prop_DNF ~ flow, RR_dnf_data)
summary(dnf_mod)
```

```
Call:
lm(formula = prop_DNF ~ flow, data = RR_dnf_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.06308 -0.04667 -0.01621  0.02705  0.09522

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.916e-02  3.810e-02   2.078    0.083 .
flow        -7.061e-06  3.273e-05  -0.216    0.836
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06784 on 6 degrees of freedom
```

```
Multiple R-squared:  0.007698,  Adjusted R-squared:  -0.1577
F-statistic: 0.04654 on 1 and 6 DF,  p-value: 0.8363
```

For every 1000 ft^3/sec increase in flow, we can expect a 0.007056 unit decrease in the proportion of DNFs.

a) The model has a negative coefficient for flow's impact on DNF proportion Does this make sense with what you saw on the graph from question 1?

It does not make sense. On the scatterplot, there is a positive trend with most of the data points.

b) Note the p-value for the flow coefficient. Is it significant?

No, the p-value for the flow coefficient is 0.836, which means there is no evidence for flow being a useful predictor in the model made using this data set.

c) Comment on the appropriateness of the model

Based on the scatterplot and the model summary, the model is not appropriate. There is one observation that is influencing the model too much for it to make sense.

Investigate Influential Observations

3. Use the augment function to add the model's residuals into the dataset

```
dnf_resid = left_join(select(RR_dnf_data, year, prop_DNF, flow), augment(dnf_mod))
```

```
Joining with `by = join_by(prop_DNF, flow)`
```

a) Do any observations have unusual leverage? standardized residual? influence?

```
dnf_resid |> filter(.hat > 2*(2/8))
```

```
# A tibble: 1 x 9
   year prop_DNF  flow .fitted  .resid  .hat .sigma .cooksd .std.resid
  <dbl>    <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl>   <dbl>      <dbl>
1  2017   0.0051  2649  0.0605 -0.0554 0.833 0.0429    10.0      -2.00
```

2017 has leverage greater than 2/8 (8 = n), making it unusual.

```
dnf_resid |> filter( abs(.std.resid) >= 2)
```

```
# A tibble: 0 x 9
# i 9 variables: year <dbl>, prop_DNF <dbl>, flow <dbl>, .fitted <dbl>,
#   .resid <dbl>, .hat <dbl>, .sigma <dbl>, .cooksd <dbl>, .std.resid <dbl>
```

All observations have standardized residuals with absolute value less than 2. There are no observations with unusual standardized residuals.

```
dnf_resid |> filter(.cooksd > 1) ## 2017 is highly influential
```

```
# A tibble: 1 x 9
   year prop_DNF  flow .fitted  .resid  .hat .sigma .cooksd .std.resid
  <dbl>    <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl>   <dbl>      <dbl>
1  2017   0.0051  2649  0.0605 -0.0554 0.833 0.0429    10.0      -2.00
```

2017 has a cook's distance of 10 (much greater than 1), making it highly influential.

Outlier Removal

4. We only want to remove outliers if we have reason to believe that outside factors may be influencing the data in a way that prevents data analysis. In this case, the outlying observation is from 2017. In 2017, the water level was dangerously high, causing the race officials to move the race start to Delaney Road (roughly 4 miles further down the river), leaving the finish in the normal place. The map below shows the race course. Do you think this change warrants removing the year from the data set? Why or why not?

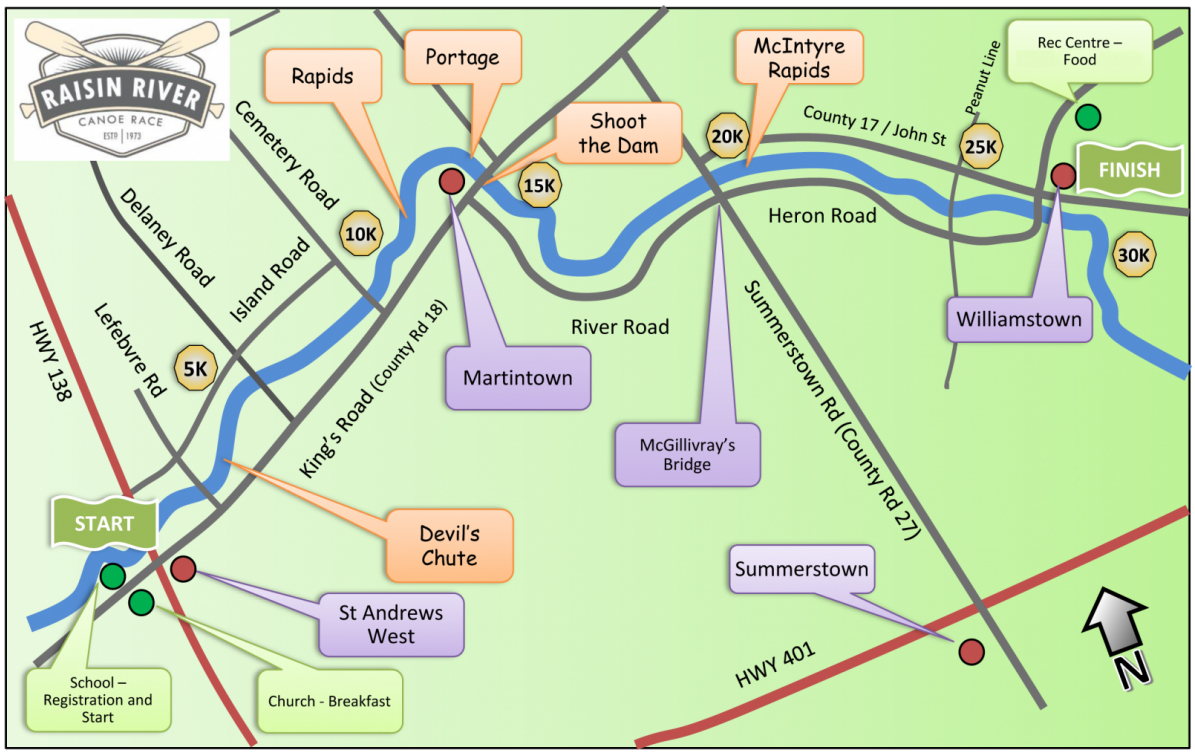


Figure 1: Raisin River Race Course

Answers may vary, but the shortening of the race course reduces the difficulty of the race notably, even with high water. This decrease in difficulty (or more importantly, effort, especially with the very fast-moving water) can account for a drop in the DNF proportion. Because of this impact, removing this observation from the data set makes sense, as all other years represent the original race course, not the shortened version.

5. Remove the influential observation from the data set.

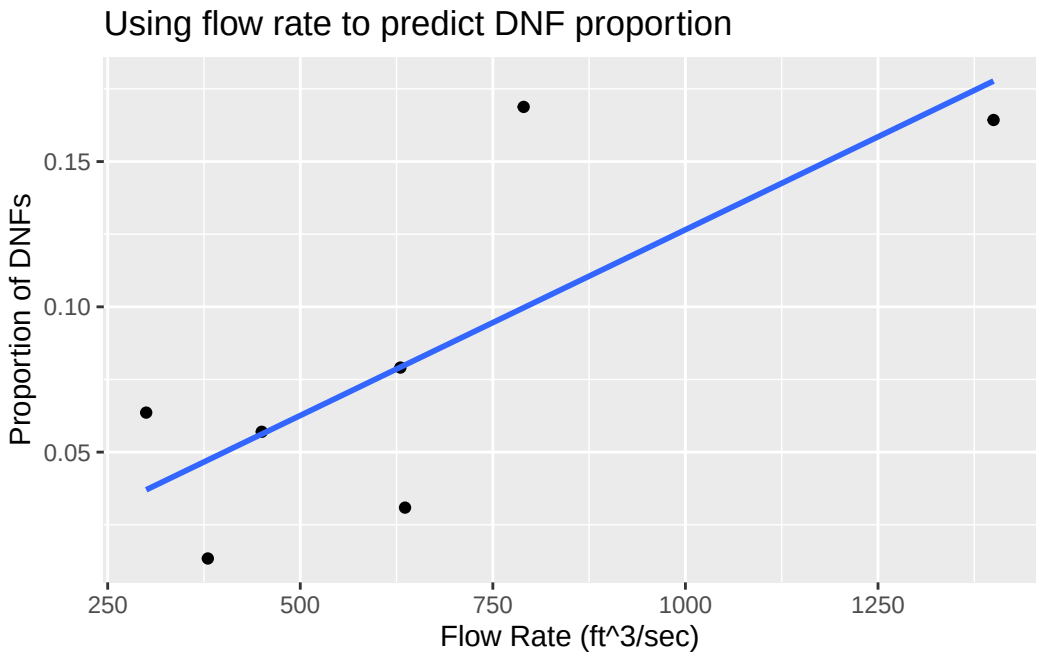
```
RR_dnf_data2 <- RR_dnf_data |> filter(year != 2017)
```

6. Regraph the scatterplot without the influential observation and add a smoother.

```
ggplot(RR_dnf_data2, aes(x = flow, y = prop_DNF)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE) +
  labs(x = "Flow Rate (ft^3/sec)",
```

```
       y = "Proportion of DNFs",
       title = "Using flow rate to predict DNF proportion")
```

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



7. Refit the model with the new version of the data set and print a summary of the model.

```
dnf_mod <- lm(prop_DNF ~ flow, RR_dnf_data2)
summary(dnf_mod)
```

```
Call:
lm(formula = prop_DNF ~ flow, data = RR_dnf_data2)

Residuals:
         1          2          3          4          5          6          7
 0.0265737 -0.0490948  0.0007913 -0.0133970 -0.0001275  0.0691113 -0.0338569

Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.338e-03  3.509e-02  -0.038   0.9710
flow         1.279e-04  4.748e-05   2.693   0.0431 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04294 on 5 degrees of freedom
Multiple R-squared:  0.592, Adjusted R-squared:  0.5104
F-statistic: 7.254 on 1 and 5 DF,  p-value: 0.04312
```

a) Meaningfully interpret the slope coefficient.

For every 100 ft^3/sec increase in flow, we can expect a 0.01279 unit increase in the proportion of DNFs.

b) If the flow level is 1100 ft^3/sec and there are 200 competitors about how many DNFs can we expect?

```
-0.001343 + (0.0001279*1100)
```

```
[1] 0.139347
```

```
200 * 0.139347
```

```
[1] 27.8694
```

If flow level is 1100 ft^3/sec and there are 200 competitors, we can expect about 27 DNFs.