DATA 606 Data Project Proposal

Julian Adames-Ng

Data Preparation

```
# load data

library(tidyverse)
library(openintro)

s_belts <- data.frame(Seatbelts)
s_belts
```

Research question

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

Do seat belt laws predict a reduction in road casualties?

Cases

What are the cases, and how many are there?

Great Britain driver deaths/seriously injured, front seat passenger deaths/seriously injured, rear-seat passenger deaths/seriously injured.

Data collection

Describe the method of data collection.

The data was collected in 1989 by A.C. Harvey, and in 2001 by J. Durbin and S. J. Koopman.

Type of study

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

This is an experimental study. We would like to see the effect of seat belt laws on rider deaths.

Data Source

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

Harvey, A.C. (1989). Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, pp. 519–523.

Durbin, J. and Koopman, S. J. (2001). Time Series Analysis by State Space Methods. Oxford University Press. http://www.ssfpack.com/dkbook/ (http://www.ssfpack.com/dkbook/)

Dependent Variable

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

There are 4 possible response variables in this study, namely the types of casualties/serious injuries between driver/front seat passengers/back seat passengers/van drivers. These are all quantitative variables.

Independent Variable

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

One independent variable is the status of the seat belt law, whether it was in effect or not. This is a qualitative variable expressed as a Boolean True/False value.

Another independent variable could be the distance driven in km. This could affect whether people wear seat belts or not. This is a quantitative value.

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.
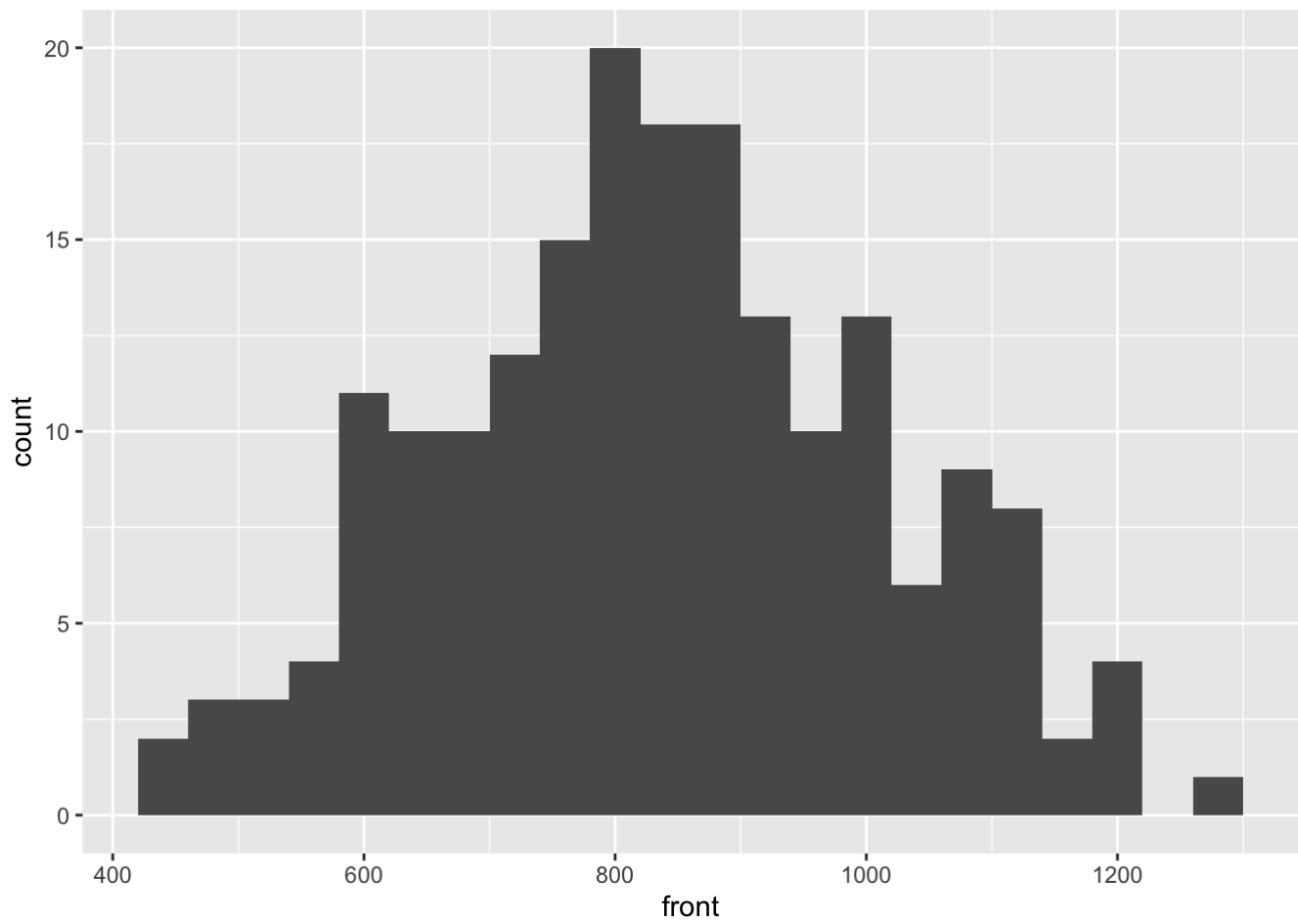
```
ggplot(s_belts, aes(x = DriversKilled)) + geom_histogram(binwidth = 40)
```
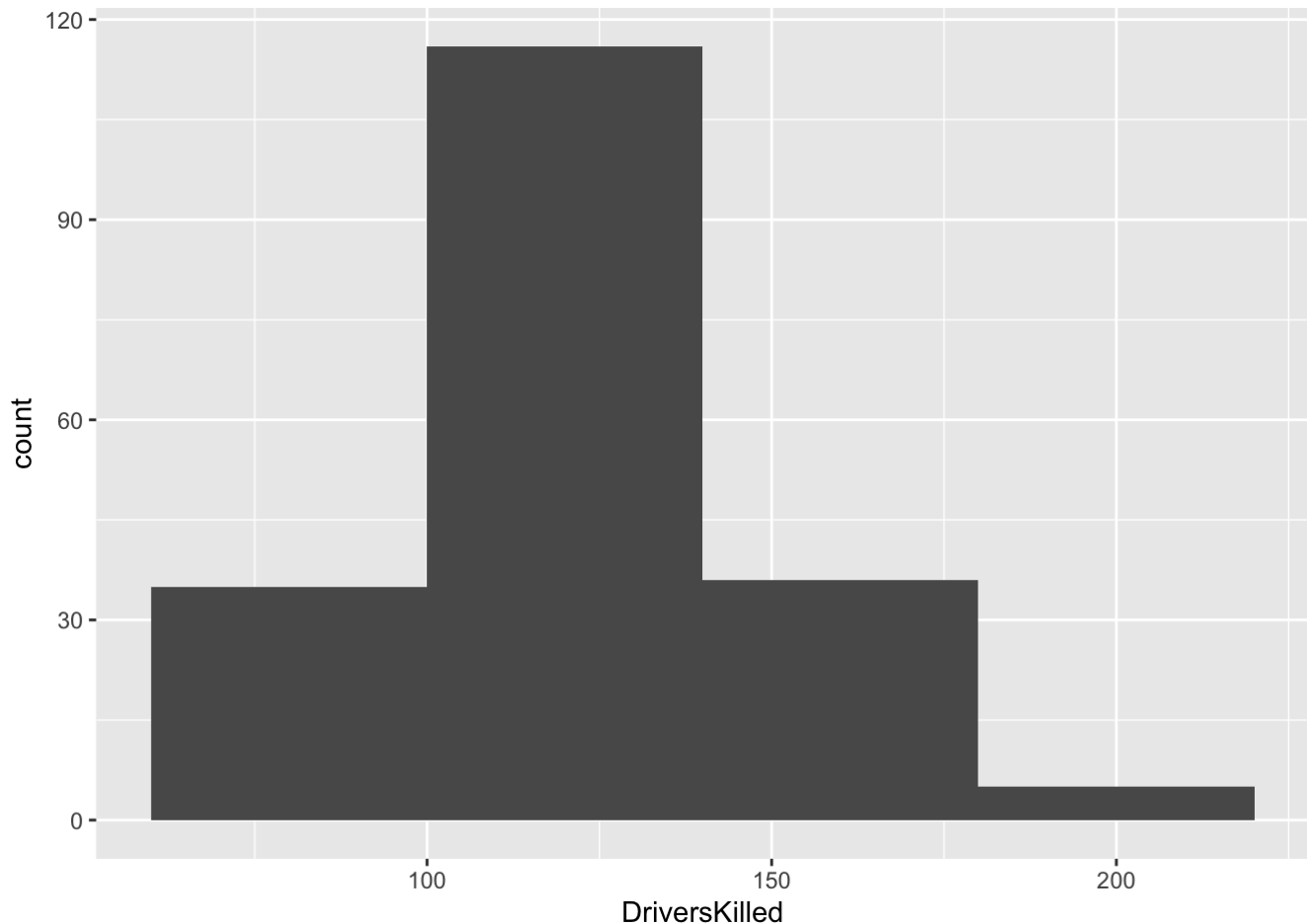
```
ggplot(s_belts, aes(x = drivers)) + geom_histogram(binwidth = 40)
```
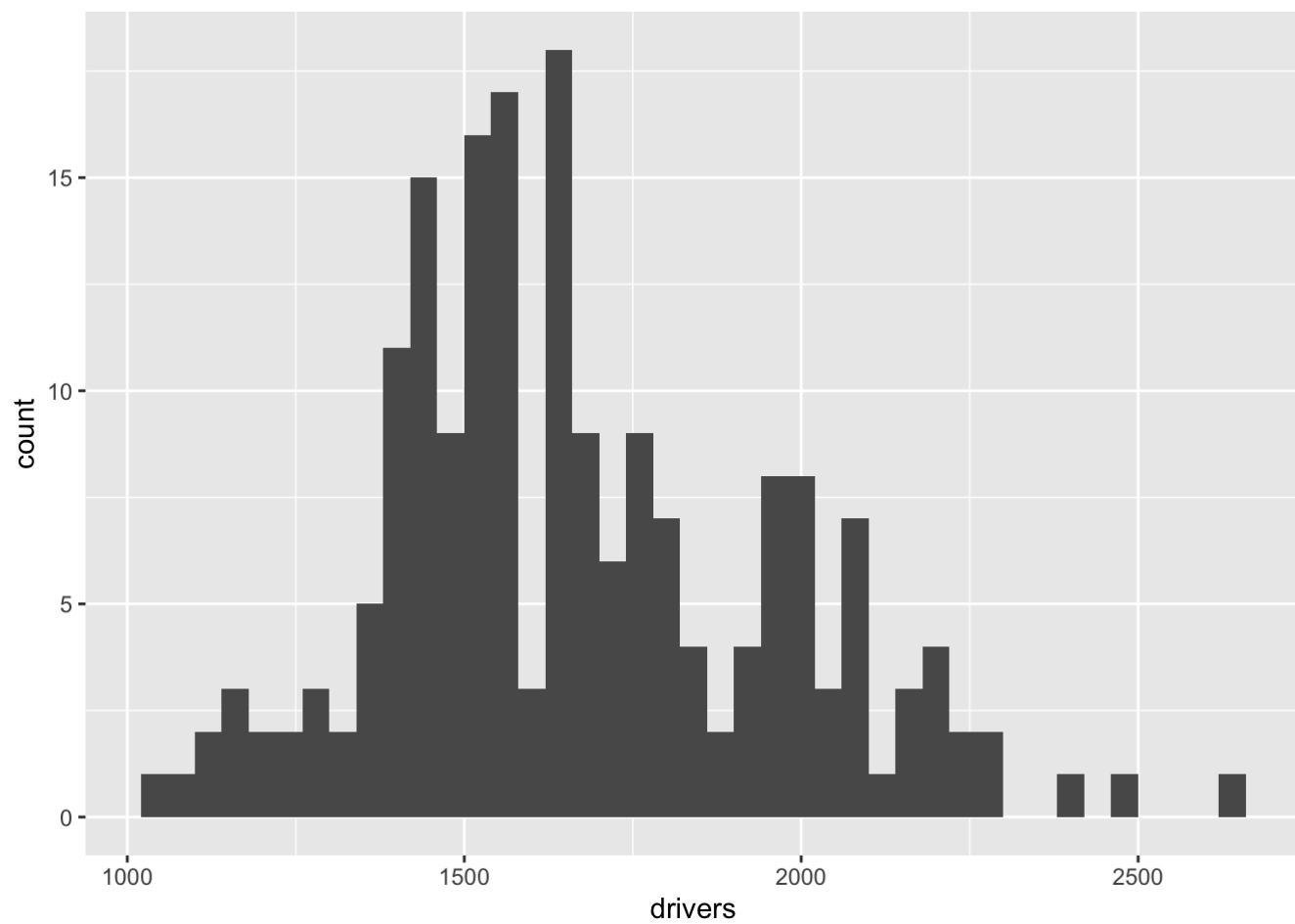


```
ggplot(s_belts, aes(x = front)) + geom_histogram(binwidth = 40)
```

```
ggplot(s_belts, aes(x = rear)) + geom_histogram(binwidth = 40)
```