PO687 assignment example

RP

17/11/2020

This document serves as a guide for your end of term assignment. Emphasis on *guide*. I will use the *beer* data set to illustrate how you can answer the question. I will offer instructions, rather than give you a set formula. What's in orange are tips from me.

You can write your assignment in a RMarkdown document, like this one. That way, you can embed R code, the output and text. This way, you wouldn't need to worry about exporting tables and graphs.

Alternatively, you can write your assignment in a Word doc and add the R code at the end of the assignment. In other word, append (copy-paste) your R script at the end of the assignment. You will not be marked on the R code, but it will help us with marking.

The rationale behind the project:

- it will test all the stats skills you acquired this term - from formulating hypotheses, to visualising relationships, running statistical analysis, presenting and interpreting the results, but also data management, such as recoding of variables, where needed;
- you have some freedom over the analysis you will run; You have to pick one of the 3 datasets available to you, and you get two pick the variables you will use in the analysis;
- rather than telling you exactly what methods to apply, you will need to think about the variables you are using and which are the appropriate statistical techniques to test the relationship(s) between the variables you chose;
- think about it as a miniature research project, but one in which you don't need a theory and literature review part. Treat this as practice for your dissertation next year (if you choose to write one).

Formulate hypotheses

1. Pick a dataset among *gss, nes* and *world*. Inspect it, have a look at the variables it contains and at the codebook. Select an outcome and a predictor variable. These will be the central elements of your assignment. Remember that the outcome variable needs to be interval, ratio or high-level ordinal - what we call a continuous variable. Feel free to recode variables, and add that information in a footnote.

Formulate the working and the null hypotheses. (15 points)

```
library(descr)
```

```
## Warning: package 'descr' was built under R version 4.0.3
```

```
library(ggplot2)
library(texreg)
```

```
## Version:  1.37.1
## Date:     2020-05-29
```

```
## Author:   Philip Leifeld (University of Essex)
##
## Please cite the JSS article in your publications -- see citation("texreg").
```

```
library(car)
```

```
## Loading required package: carData
```

```
setwd("C:/Users/poppr/Dropbox/kent/PO687/week 3/seminar")
d <- read.csv('beer.csv', stringsAsFactors = T)
names(d)
```

```
## [1] "abv"        "ibu"        "id"         "name"       "style"
## [6] "brewery_id" "ounces"
```

```
str(d)
```

```
## 'data.frame':    2410 obs. of  7 variables:
##  $ abv       : num  0.05 0.066 0.071 0.09 0.075 0.077 0.045 0.065 0.055 0.086 ...
##  $ ibu       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ id        : int  1436 2265 2264 2263 2262 2261 2260 2259 2258 2131 ...
##  $ name      : Factor w/ 2305 levels "#001 Golden Amber Lager",..: 1638 577 1705 1842 1819 268 1160 
##  $ style     : Factor w/ 100 levels "","Abbey Single Ale",..: 19 18 16 12 16 80 18 22 18 12 ...
##  $ brewery_id: int  408 177 177 177 177 177 177 177 177 177 ...
##  $ ounces    : num  12 12 12 12 12 12 12 12 12 12 ...
```

Possible answer For this assignment, I chose to use the 'beer' data set. After a close inspection of the data set and of the coodebok, I decided to inspect the relationship between Alcohol by volume and International Bitterness Units. In other words, I want to see whether bitterness is a good predictor for alcohol percentage in beer.

The outcome variable is *abv* Alcohol by volume. The predictor variable is *ibu* International Bitterness Units scale

H1: Beers with a higher IBU tend to have a higher ABV * H1 could also could also sound like this: Bitter beers tend to be more alcoholic * Bitter beers have a higher alcohol content.

H0: There is no relationship between how bitterness and alcohol content in beer

Univariate statistics and visualisations

2. Describe the two variables. Create appropriate visualisations for each variable, accompanied by the appropriate descriptive statistics (hint: it all depends on the level of measurement). (15 points)

```
table(d$abv)
```

```
##
## 0.001 0.027 0.028 0.032 0.034 0.035 0.037 0.038 0.039  0.04 0.041 0.042 0.043
##     1     2     1     3     1     6     4     6    15    38     9    38    12
## 0.044 0.045 0.046 0.047 0.048 0.049  0.05 0.051 0.052 0.053 0.054 0.055 0.056
##    16    89    32    57    72    59   215    62   107    60    50   158    66
## 0.057 0.058 0.059  0.06 0.061 0.062 0.063 0.064 0.065 0.066 0.067 0.068 0.069
##    52    66    34   125    21    59    38    21   123    20    25    52    32
##  0.07 0.071 0.072 0.073 0.074 0.075 0.076 0.077 0.078 0.079  0.08 0.081 0.082
##    92    18    37    18     8    43     5    13    10     7    57     4    22
## 0.083 0.084 0.085 0.086 0.087 0.088 0.089  0.09 0.091 0.092 0.093 0.094 0.095
##     8     3    26     4    10     6     2    24     3    13     5     1     9
## 0.096 0.097 0.098 0.099   0.1 0.104  0.12 0.125 0.128
##     5     5     3    35     1     1     1     1     1
```

The outcome variable, *abv*, is *Alcohol by volume*. It is a ratio variable, therefore appropriate to use as an outcome variable.

```
mean(d$ibu, na.rm=T)
```

```
## [1] 42.71317
```

```
median(d$ibu, na.rm=T)
```

```
## [1] 35
```

```
range(d$ibu, na.rm=T)
```

```
## [1]   4 138
```

```
sd(d$ibu, na.rm=T)
```

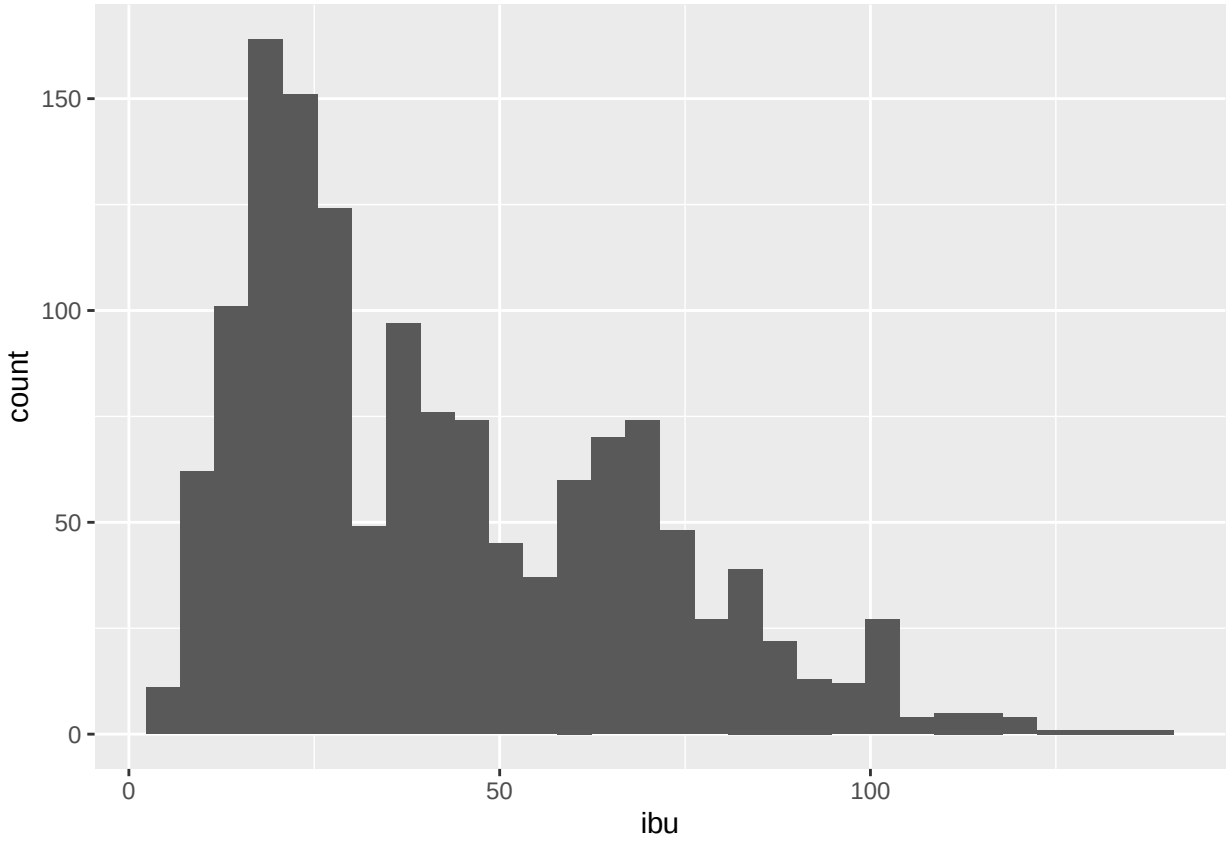
```
## [1] 25.95407
```

The predictor variable is *ibu*, *International Bitterness Units scale*. The level of measurement of this variable is interval. The mean of the variable is 42.71, and the median is 35. Its minimum and maximum values are 4 and 138. The standard deviation is 25.95.

```
ggplot(d, aes(ibu)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 1005 rows containing non-finite values (stat_bin).
```

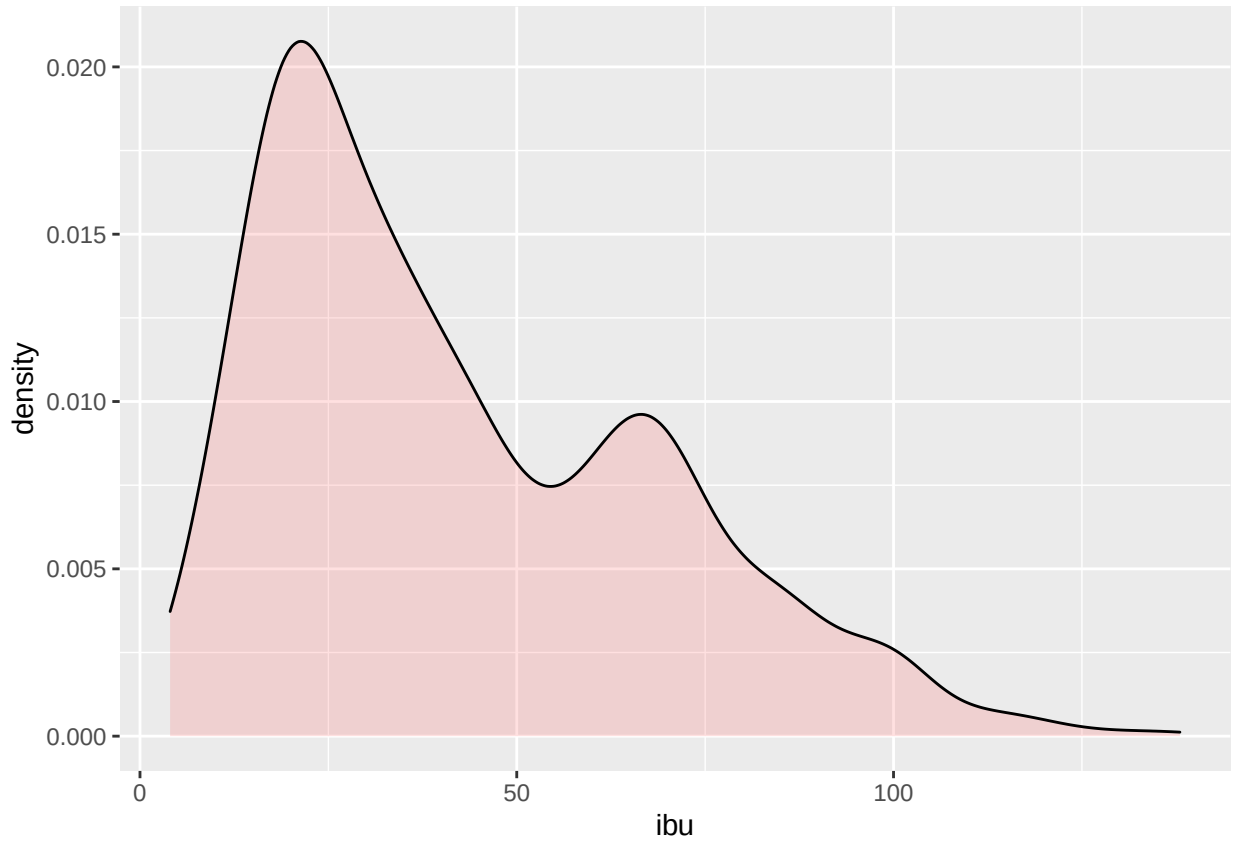


Make this graph prettier, by changing the label, adding a title, using a different theme for the plot.

Or you can create one with a density line.

```
ggplot(d, aes(ibu)) +
  geom_density(alpha=.2, fill="#FF6666")
```

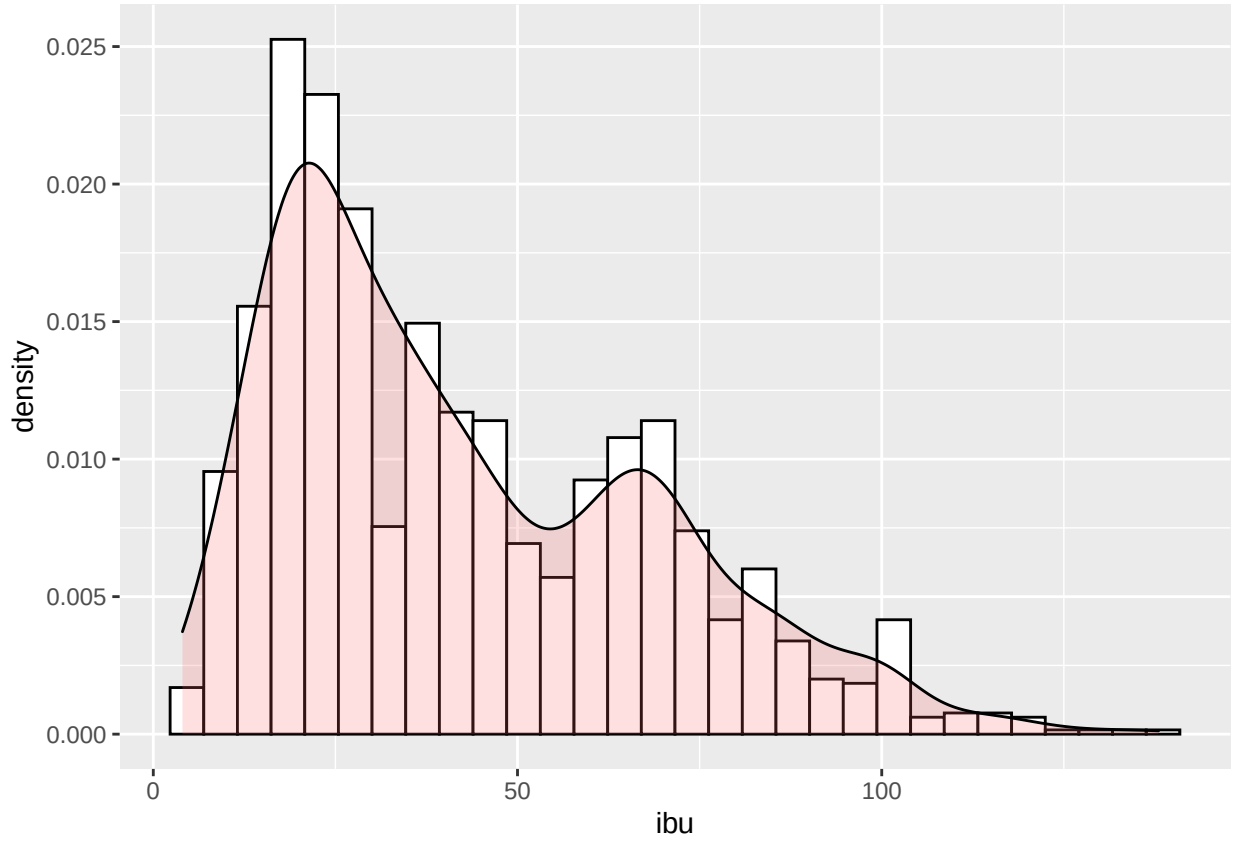
```
## Warning: Removed 1005 rows containing non-finite values (stat_density).
```



Or even better, one that combines the two.

```
ggplot(d, aes(ibu)) +
  geom_histogram(aes(y=..density..),
                 colour="black", fill="white") +
  geom_density(alpha=.2, fill="#FF6666")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 1005 rows containing non-finite values (stat_bin).
## Warning: Removed 1005 rows containing non-finite values (stat_density).
```



Based on the histogram, we can see that the IBU variable has a positive skews, with most beers in the data set having beers with lower IBU values. There are some extreme values toward the higher end of the scale.

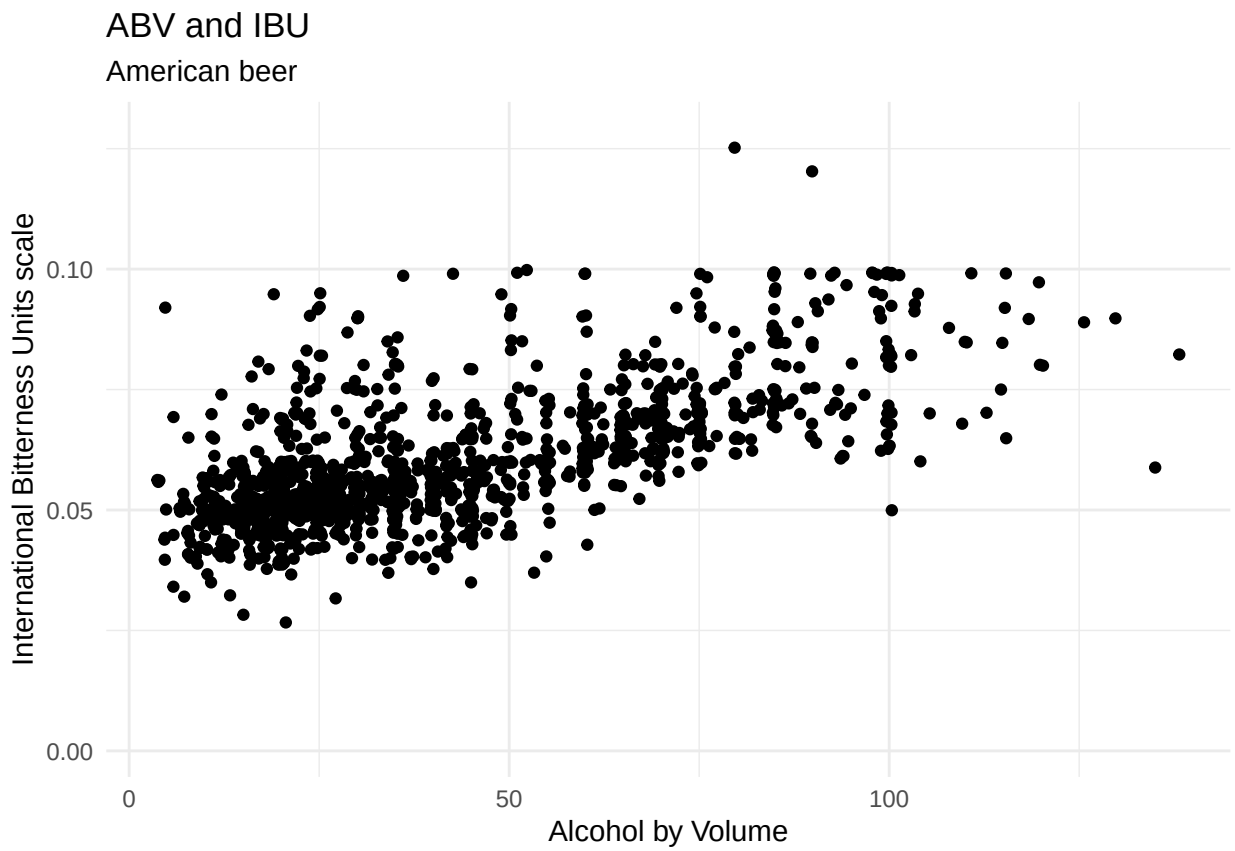
Describe the outcome variable in a similar way. Pay attention to the level of measuremnt, that will dictate what kind of graoh to use.

Visualise a bivariate relationship

3. Thinking about the type of variable you selected, create a graph that will illustrate the relationship between your dependent and independent variables. Remember that visualisations have to be nice to look at, represent the data truthfully, be clear and informative. In other words, do not forget to add titles, labels and so on. (15 points)

```
ggplot(d, aes(ibu, abv)) +
  geom_point(position='jitter') +
  labs(title = 'ABV and IBU',
       subtitle = 'American beer',
       x = 'Alcohol by Volume',
       y = 'International Bitterness Units scale') +
  theme_minimal()
```

```
## Warning: Removed 1005 rows containing missing values (geom_point).
```



The title and subtitiles are silly, you can find better ones.

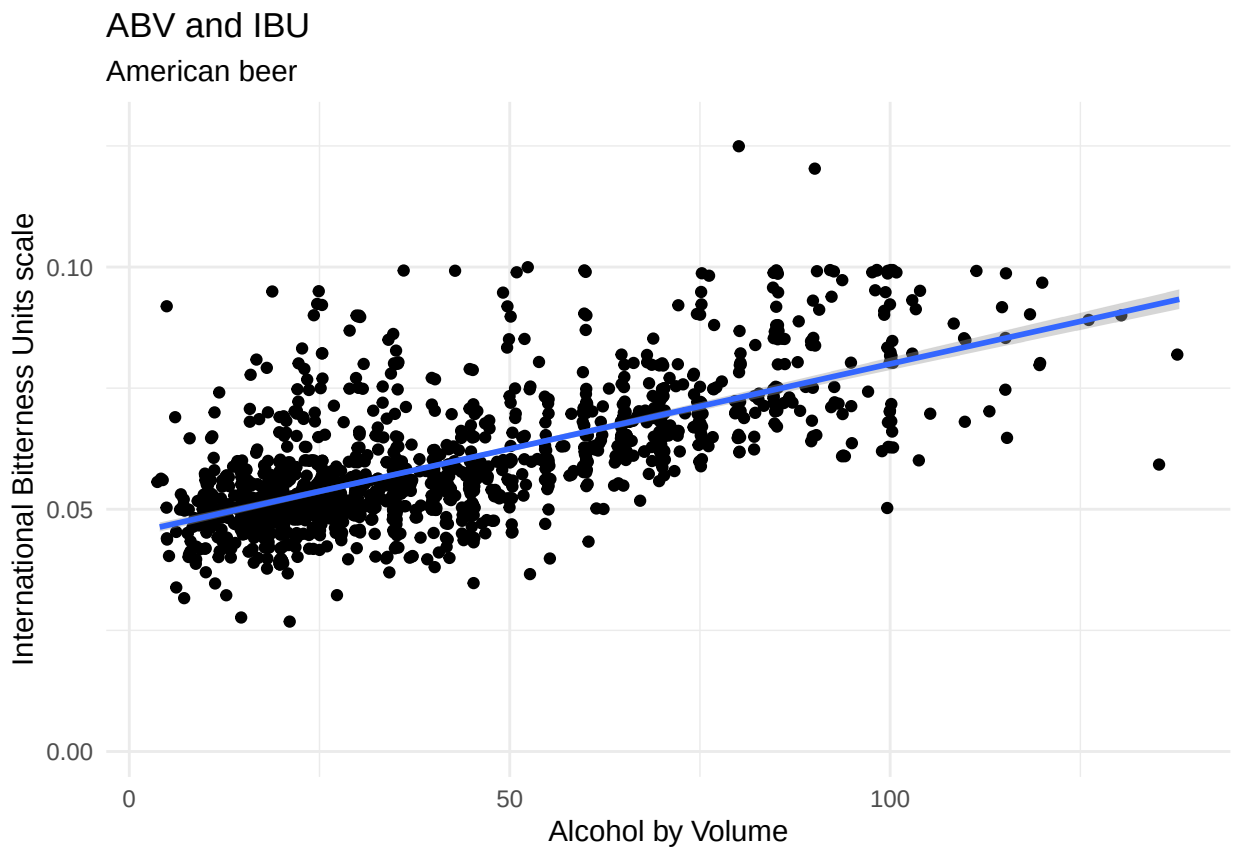
Because in this situation makes sense, you could add a regression line.

```
ggplot(d, aes(ibu, abv)) +
  geom_point(position='jitter') +
    geom_smooth(method = 'lm') +
    labs(title = 'ABV and IBU',
       subtitle = 'American beer',
       x = 'Alcohol by Volume',
       y = 'International Bitterness Units scale') +
  theme_minimal()
```

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

```
## Warning: Removed 1005 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1005 rows containing missing values (geom_point).
```



The graph above illustrates the relationship between the outcome variable *Alcohol by Volume* and the predictor *International Bitterness Units scale*. There is a pattern in the data, which suggests a positive relationship. As ABV goes up, so does IBU. Data seems to be clustered around the lower values of ABV and of IBU. The regression line, which is the line of best fit, has a positive slope, which suggests that there is indeed a positive correlation between the two variables. Judging by the slope of the line how steep it is the relationship seems to be moderately strong.

Hypothesis testing with a t-test or a non-parametric test

4. Test the hypothesis you formulated in *Step 1* using a t-test or a non-parametric test, depending on which one is appropriate (hint: remember it depends on whether the variable is normally distributed or not). Report the test statistics, and its associated p-value. Use the .05 cut off point for statistical significance and interpret the results. (15 points)

Here you need to think about the level of measurement of the two variables, that will dictate the type of test you need to run.

Because both variables are interval level (or continuous) to test the hypothesis, I will run a correlation test. I set the cutoff point for statistiocal significance at .05.

```
cor.test(d$abv, d$ibu)
```

```
##
##  Pearson's product-moment correlation
##
## data:  d$abv and d$ibu
## t = 33.863, df = 1403, p-value < 2.2e-16
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6407982 0.6984238
## sample estimates:
##       cor
## 0.6706215
```

The Pearson R coefficient is 0.67, with p-value < 2.2e-16. Because the p-value is lower than .05, we can reject the null hypothesis which says that there is no relationship between ABV and IBU. The value of the correlation coefficient suggests that there is a moderately strong relationship between bitterness and alcohol content in American beer.

Simple as that.

Bivariate regression

5. Test the hypothesis you formulated in *Step 1* using a regression model. Present the regression results in a table and interpret them. Use the .05 cut off point for statistical significance.

(15 points)

```
m <- lm(abv ~ ibu, d)
summary(m)
```

```
##
## Call:
## lm(formula = abv ~ ibu, data = d)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.033288 -0.005946 -0.001595  0.004022  0.052006
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.493e-02  5.177e-04   86.79   <2e-16 ***
## ibu         3.508e-04  1.036e-05   33.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01007 on 1403 degrees of freedom
##   (1005 observations deleted due to missingness)
## Multiple R-squared:  0.4497, Adjusted R-squared:  0.4493
## F-statistic:  1147 on 1 and 1403 DF,  p-value: < 2.2e-16
```

if you write your assignment in RMarkdown, you can use the screenreg() function from the texreg package. it will create a nice table, like the one below. If you write your assignment in word, then use the htmlreg(), as you practiced during the seminars.

```
screenreg(m)
```

```
##
## ========================
##              Model 1
## ------------------------
```

```
## (Intercept)     0.04 ***
##                (0.00)
## ibu             0.00 ***
##                (0.00)
## ------------------------
## R^2             0.45
## Adj. R^2        0.45
## Num. obs.    1405
## ========================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

The R^2 and Adjusted R^2 have the value of 0.45, meaning that the model explains 45% of variation in the outcome variable, ABV. The residuals are symmetrical, which indicates that our model has a significant p-value for the F-test.

The value of the intercept is 4.493e-02. The regression coefficient is 3.508e-04. When a beer's IBV increases with 1 unit, the alcohol content of the beer increases by 3.508e-04 points. The regression coefficient is statistically significant at the .05 level. This means that we can reject the null hypothesis. In other words, the regression model shows that bitter beers tend to have higher levels of alcohol. However, the effect is rather small.

This model is not great as an example, becasue the values in our variable are 0.something. Ideally, I would recode this variable by multiplying it by 100, because as it is now, the values are percentages. That's why we have the really tiny regression coefficients.

Multiple regression

6. Expand on the relationship you tested above, by choosing another two variables that could improve your model. Feel free to recode variables.

I want to see whether the style of beer can predict the ABV. Looking at the *style* variable, I see that it needs recoding, because it has too many values. I will create a new variable called 'stout', which will take the value of 1 for all stouts, and 0 for everything else.

The other predictor I will look at is *ounces*, the volume in which the beer is sold. This actually means the size of the can or bottle.

```
table(d$style)
```

```
##
##                                          Abbey Single Ale
##                                 5                        2
##                           Altbier      American Adjunct Lager
##                                13                       18
##        American Amber / Red Ale    American Amber / Red Lager
##                               133                       29
##            American Barleywine           American Black Ale
##                                3                       36
##            American Blonde Ale           American Brown Ale
##                               108                       70
##         American Dark Wheat Ale American Double / Imperial IPA
##                                7                      105
##  American Double / Imperial Pilsner American Double / Imperial Stout
##                                2                        9
##       American India Pale Lager                  American IPA
##                                3                      424
##            American Malt Liquor       American Pale Ale (APA)
```

```
##                                1                                245
##                American Pale Lager            American Pale Wheat Ale
##                                 39                                97
##                   American Pilsner                  American Porter
##                                 25                                68
##                     American Stout               American Strong Ale
##                                 39                                14
##                 American White IPA                 American Wild Ale
##                                 11                                 6
##                      Baltic Porter                  Belgian Dark Ale
##                                  6                                11
##                      Belgian IPA                  Belgian Pale Ale
##                                 18                                24
##            Belgian Strong Dark Ale            Belgian Strong Pale Ale
##                                  6                                 7
##                 Berliner Weissbier                  BiÃ¨re de Garde
##                                 11                                 7
##                               Bock                           Braggot
##                                  7                                 1
##     California Common / Steam Beer                       Chile Beer
##                                  6                                 3
##                              Cider                        Cream Ale
##                                 37                                29
##                     Czech Pilsener                       Doppelbock
##                                 28                                 7
##          Dortmunder / Export Lager                           Dubbel
##                                  6                                 5
##                       Dunkelweizen                English Barleywine
##                                  4                                 3
##                     English Bitter                 English Brown Ale
##                                  3                                18
##               English Dark Mild Ale        English India Pale Ale (IPA)
##                                  6                                13
##                    English Pale Ale               English Pale Mild Ale
##                                 12                                 3
##                      English Stout                 English Strong Ale
##                                  2                                 4
##                     Euro Dark Lager                    Euro Pale Lager
##                                  5                                 2
## Extra Special / Strong Bitter (ESB)                 Flanders Oud Bruin
##                                 20                                 1
##                     Flanders Red Ale               Foreign / Export Stout
##                                  1                                 6
##              Fruit / Vegetable Beer                    German Pilsener
##                                 49                                36
##                               Gose                          Grisette
##                                 10                                 1
##                         Hefeweizen                 Herbed / Spiced Beer
##                                 40                                 9
##                      Irish Dry Stout                       Irish Red Ale
##                                  5                                12
##                             KÃ¶lsch           Keller Bier / Zwickel Bier
##                                 42                                 3
##                       Kristalweizen                      Light Lager
```

```
##                                   1                                  12
##                      Low Alcohol Beer                MÃ¤rzen / Oktoberfest
##                                   1                                  30
##                Maibock / Helles Bock                                Mead
##                                   5                                   5
##                   Milk / Sweet Stout                  Munich Dunkel Lager
##                                  10                                   4
##                  Munich Helles Lager                       Oatmeal Stout
##                                  20                                  18
##                               Old Ale                               Other
##                                   2                                   1
##                           Pumpkin Ale                     Quadrupel (Quad)
##                                  23                                   4
##                               Radler                            Rauchbier
##                                   3                                   2
##                            Roggenbier                 Russian Imperial Stout
##                                   2                                  11
##                              Rye Beer                 Saison / Farmhouse Ale
##                                  18                                  52
##                            Schwarzbier                 Scotch Ale / Wee Heavy
##                                   9                                  15
##                           Scottish Ale                               Shandy
##                                  19                                   3
##                            Smoked Beer                               Tripel
##                                   1                                  11
##                           Vienna Lager                           Wheat Ale
##                                  20                                   1
##                          Winter Warmer                             Witbier
##                                  15                                  51
```

```
d$stout <- recode(d$style, "'Oatmeal Stout' = 1;
                  'Milk / Sweet Stout'=1;
                  'Irish Dry Stout'=1;
                  'Foreign / Export Stout'=1;
                  'English Stout' = 1;
                  'Russian Imperial Stout'=1;
                  'American Stout'=1;
                  'American Double / Imperial Stout'=1;
                  else=0")

table(d$stout)
```

```
##
##    0    1
## 2310  100
```

check the new variable

```
class(d$stout)
```

```
## [1] "factor"
```

6a. Create hypotheses for each new variable (and your outcome variable). (5 points)

H1: Stouts tend to have a higher ABV compared to the rest

H0:

H1: beer in smaller cans/bottles tends to have a higher ABV

H0:

6b. Present univariate analysis on the new variables (descriptive statistics and visualisations). (5 points)

I am showing this above

Before you include them in the model, you need to make sure that the new variables make the cut. There are a few aspects that you'll learn how to address in Lecture 10, when we talk about linear regression assumptions.

6c. Run a regression model that includes the new variables. Present the regression results in a table and interpret them. Use the .05 cut off point for statistical significance. (10 points)

```
mv <- lm(abv ~ ibu + stout + ounces, d)
summary(mv)
```

```
##
## Call:
## lm(formula = abv ~ ibu + stout + ounces, data = d)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.032193 -0.005898 -0.001425  0.003837  0.051434
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.965e-02  1.651e-03  24.020  < 2e-16 ***
## ibu         3.481e-04  1.024e-05  33.998  < 2e-16 ***
## stout1      7.459e-03  1.421e-03   5.249 1.77e-07 ***
## ounces      3.792e-04  1.181e-04   3.210  0.00136 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00994 on 1401 degrees of freedom
##   (1005 observations deleted due to missingness)
## Multiple R-squared:  0.4651, Adjusted R-squared:  0.4639
## F-statistic:   406 on 3 and 1401 DF,  p-value: < 2.2e-16
```

```
screenreg(mv)
```

```
##
## ========================
##              Model 1
## ------------------------
## (Intercept)     0.04 ***
##                (0.00)
## ibu             0.00 ***
##                (0.00)
## stout1          0.01 ***
##                (0.00)
## ounces          0.00 **
##                (0.00)
## ------------------------
```

```
## R^2             0.47
## Adj. R^2        0.46
## Num. obs.    1405
## ========================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

I am showing this above. All you need to remember: 'while controlling for'.

6d. Compare the new regression model to the model from *Step 5*, using the appropriate statistical test. Report the results and interpret them. Is the second regression model more informative? (5 points)

```
anova(m, mv)
```

```
## Analysis of Variance Table
##
## Model 1: abv ~ ibu
## Model 2: abv ~ ibu + stout + ounces
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   1403 0.14240
## 2   1401 0.13843  2 0.0039689 20.084 2.514e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You will learn more about this in week 10, but the idea is that we are comparing the F tests of our regression models, uisng an ANOVA test. If the p-value is siginificant, it means that the second model is a better fit to the data. Which basically means that it helps explain the variation in the outcome variable bettwen than the first model.

Looking at the Adjusted R^2 values, the second model explains more variation in ABV.

The results of the ANOVA test indicates that model 2 is a better fit to the data, because the p-value is $<.05$.