HW 1

ZOPE

2022-04-01

1. The main difference has to do with whether the data is labelled. In supervised learning, the data and output are labelled. Therefore, the goal of supervised learning is to determine a function which correctly represents the relationship between the output and other data. With unsupervised learning, the data is not labelled. Therefore, the goal of unsupervised learning is to discover the structure of the unlabelled data.
2. A regression model has to do with the type of output. A regression model involves a quantitative output. This has to do with numerical values and can refer to quantities such as prices, miles, etc. A classification model involves a qualitative output. This output is nonnumerical in nature and can refer to qualities such as health status (alive/not alive), color of car, etc.
3. Two commonly used metrics for regression ML problems are mean squared error (MSE) and mean absolute regression (MAE). Two commonly used metrics for classification ML problems are Accuracy and F1-score.
4. Descriptive: Here we are using a model to visually emphasize a pattern within our data. Predictive: Here we are using a model to predict an outcome (Y) with as little error as
possible. Inferential: Here we are using a model to test theories. These theories allow us to ask questions such as: What features of the models are significant? Can our data be
generalized to the broader population? etc.
5. A mechanistic model assumes that the model is parametric in nature. From there, we can add more parameters based off of the nature of the data. An empirical model relies more on real-world data instead of theory. Here, we look at the data and have much more flexibility in creating a model in comparison to a mechanistic model. Both of these models run the risk of overfitting,
a phenomenon where the model is to closely aligned to a particular set of data points.

In general, an empirical model is easier to understand. This is because the nature of the model will always be parametric.

A model that will be less flexible (empirical model) will reduce bias and therefore provide more interpretability. However, this will also lead to a natural increase in variance (spread of data). On the other hand, a model that is less flexible (mechanistic) will increase bias and therefore decrease interpretability. However, this will also lead to a natural decrease in variance (spread of data). This is the natural tradeoff we must deal with when choosing models.

6. The first statement is predictive in nature. This is because you are predicting an output (vote or no vote) based off of different variables. The second statement is inferential. Here, you are asking a question about a potential variable and whether it is relevant to the model.

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.4     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.3     ✓ stringr 1.4.0
## ✓ readr   2.0.1     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidymodels)
```

```
## Registered S3 method overwritten by 'tune':
##   method                   from
##   required_pkgs.model_spec parsnip
```

```
## ── Attaching packages ────────────────────────────────────── tidymodels 0.1.3 ──
```

```
## ✓ broom        0.7.9      ✓ rsample      0.1.0
## ✓ dials        0.0.10     ✓ tune         0.1.6
## ✓ infer        1.0.0      ✓ workflows    0.2.3
## ✓ modeldata    0.1.1      ✓ workflowsets 0.1.0
## ✓ parsnip      0.1.7      ✓ yardstick    0.0.8
## ✓ recipes      0.1.16
```

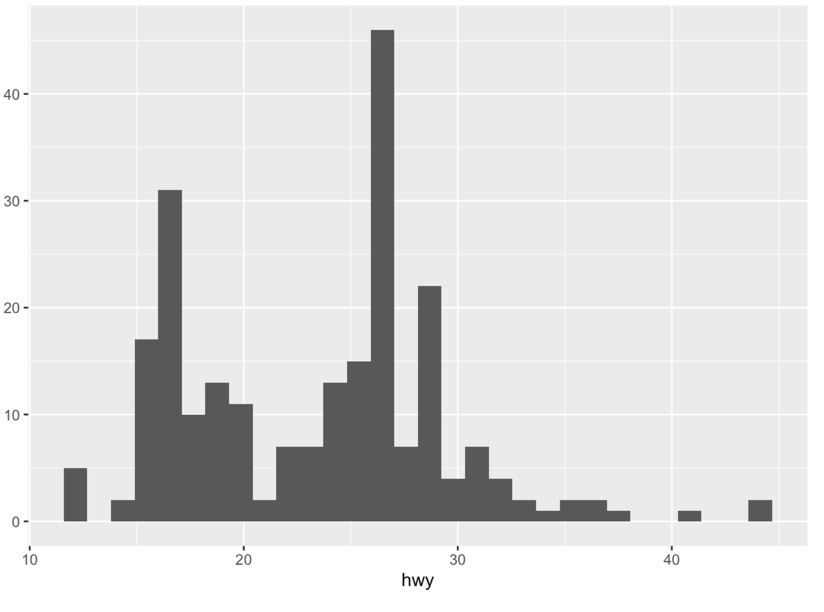
```
## ── Conflicts ───────────────────────────────────────── tidymodels_conflicts() ──
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
## • Use tidymodels_prefer() to resolve common conflicts.
```

```
library(ISLR)
mpg
```

```
## # A tibble: 234 × 11
##    manufacturer model      displ  year   cyl trans drv     cty   hwy fl    class
##    <chr>        <chr>      <dbl> <int> <int> <chr> <chr> <int> <int> <chr> <chr>
##  1 audi         a4           1.8  1999     4 auto… f        18    29 p     comp…
##  2 audi         a4           1.8  1999     4 manu… f        21    29 p     comp…
##  3 audi         a4           2    2008     4 manu… f        20    31 p     comp…
##  4 audi         a4           2    2008     4 auto… f        21    30 p     comp…
##  5 audi         a4           2.8  1999     6 auto… f        16    26 p     comp…
##  6 audi         a4           2.8  1999     6 manu… f        18    26 p     comp…
##  7 audi         a4           3.1  2008     6 auto… f        18    27 p     comp…
##  8 audi         a4 quattro   1.8  1999     4 manu… 4        18    26 p     comp…
##  9 audi         a4 quattro   1.8  1999     4 auto… 4        16    25 p     comp…
## 10 audi         a4 quattro   2    2008     4 manu… 4        20    28 p     comp…
## # … with 224 more rows
```

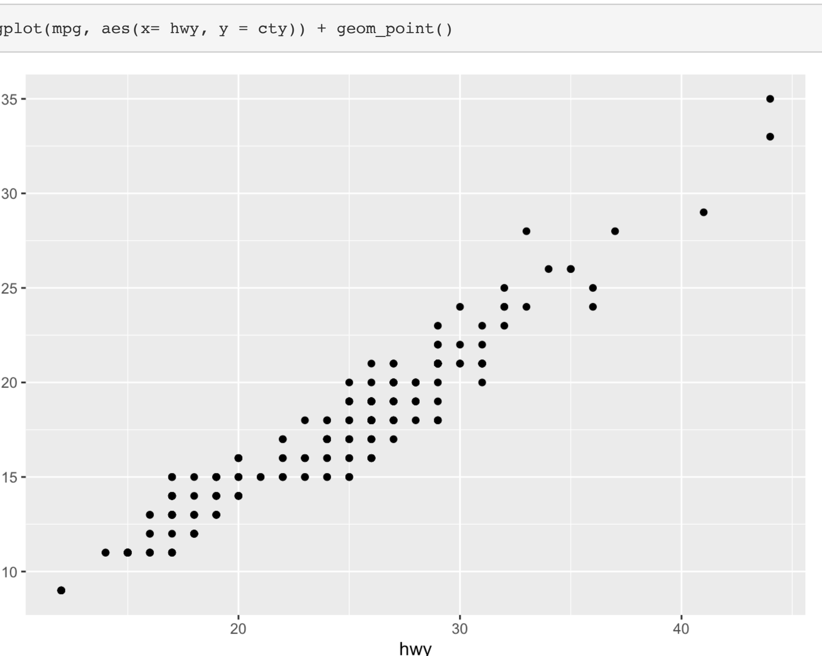
```
ggplot(data=mpg, aes(hwy)) +
geom_histogram()
```

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



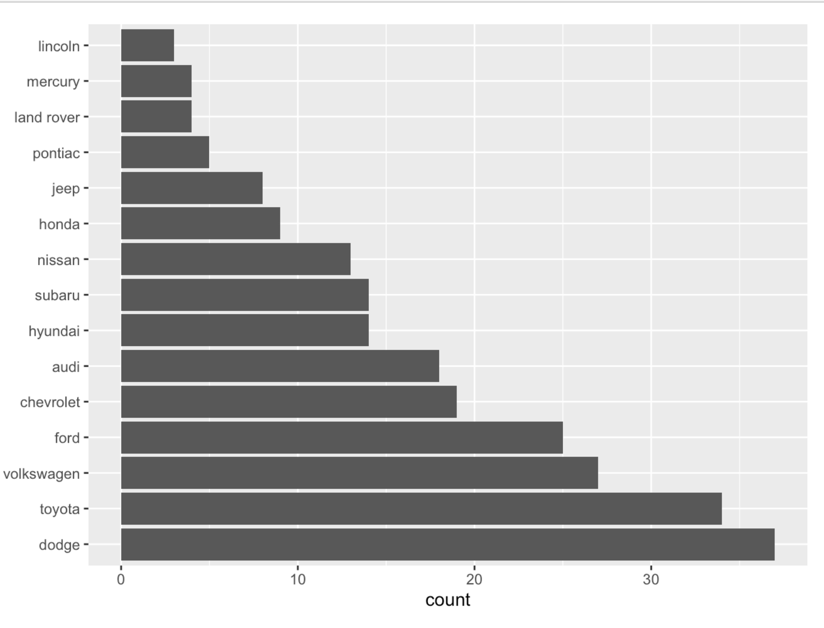
The histogram tells us that we see the highest amount of vehicles lie between 25-30 miles per gallon range and 15-20 miles per gallon range. It also indicates that we see the fewest amount of vehicles lie below 15 miles per gallon and above 35 miles per gallon.

```
ggplot(mpg, aes(x= hwy, y = cty)) + geom_point()
```



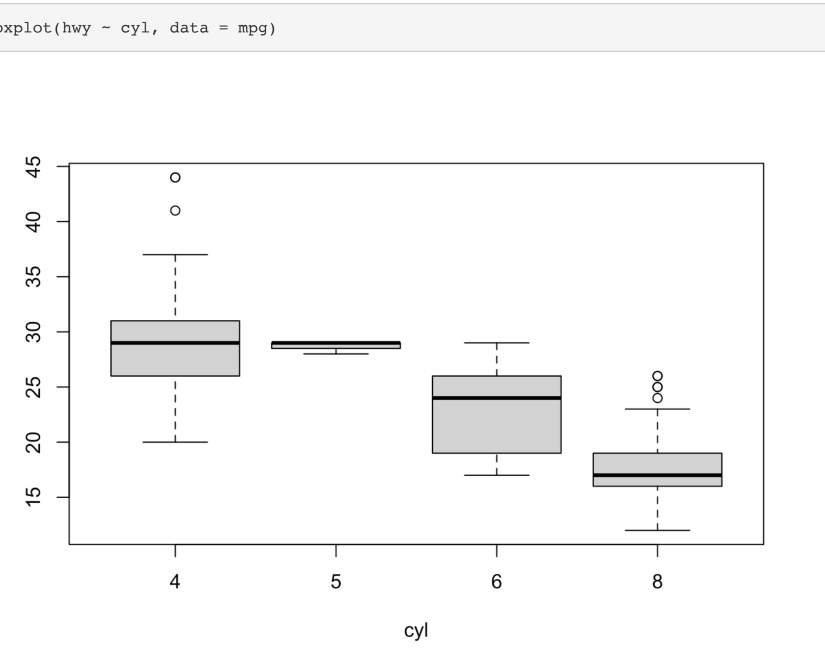
The scatterplot indicates a positive and linear relationship between hwy and cty. This indicates that for each highway mile per gallon increase, we should see a roughly constant increase in the corresponding city miles per gallon, all else being held equal.

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



Dodge produced the most cars and Lincoln produced the least.

```
boxplot(hwy ~ cyl, data = mpg)
```



We see a negative relationship between the number of highway miles per gallon and the number of cylinders present. All else being held equal, an increase in the number of cylinders will lead to a decrease in the median number of highway miles per gallon.

```
df = subset(mpg, select = -c(manufacturer, displ, drv, model, trans, fl, class) )
#install.packages("corrplot")
mpg_numeric <- round(cor(df),2)
lower.tri(df, diag = FALSE)
```

```
##         [,1]  [,2]  [,3]  [,4]
##   [1,] FALSE FALSE FALSE FALSE
##   [2,]  TRUE FALSE FALSE FALSE
##   [3,]  TRUE  TRUE FALSE FALSE
##   [4,]  TRUE  TRUE  TRUE FALSE
##   [5,]  TRUE  TRUE  TRUE  TRUE
##   [6,]  TRUE  TRUE  TRUE  TRUE
##   [7,]  TRUE  TRUE  TRUE  TRUE
##   [8,]  TRUE  TRUE  TRUE  TRUE
##   [9,]  TRUE  TRUE  TRUE  TRUE
##  [10,]  TRUE  TRUE  TRUE  TRUE
##  [11,]  TRUE  TRUE  TRUE  TRUE
##  [12,]  TRUE  TRUE  TRUE  TRUE
##  [13,]  TRUE  TRUE  TRUE  TRUE
##  [14,]  TRUE  TRUE  TRUE  TRUE
##  [15,]  TRUE  TRUE  TRUE  TRUE
##  [16,]  TRUE  TRUE  TRUE  TRUE
##  [17,]  TRUE  TRUE  TRUE  TRUE
##  [18,]  TRUE  TRUE  TRUE  TRUE
##  [19,]  TRUE  TRUE  TRUE  TRUE
##  [20,]  TRUE  TRUE  TRUE  TRUE
##  [21,]  TRUE  TRUE  TRUE  TRUE
##  [22,]  TRUE  TRUE  TRUE  TRUE
##  [23,]  TRUE  TRUE  TRUE  TRUE
##  [24,]  TRUE  TRUE  TRUE  TRUE
##  [25,]  TRUE  TRUE  TRUE  TRUE
##  [26,]  TRUE  TRUE  TRUE  TRUE
##  [27,]  TRUE  TRUE  TRUE  TRUE
##  [28,]  TRUE  TRUE  TRUE  TRUE
##  [29,]  TRUE  TRUE  TRUE  TRUE
##  [30,]  TRUE  TRUE  TRUE  TRUE
##  [31,]  TRUE  TRUE  TRUE  TRUE
##  [32,]  TRUE  TRUE  TRUE  TRUE
##  [33,]  TRUE  TRUE  TRUE  TRUE
##  [34,]  TRUE  TRUE  TRUE  TRUE
##  [35,]  TRUE  TRUE  TRUE  TRUE
##  [36,]  TRUE  TRUE  TRUE  TRUE
##  [37,]  TRUE  TRUE  TRUE  TRUE
##  [38,]  TRUE  TRUE  TRUE  TRUE
##  [39,]  TRUE  TRUE  TRUE  TRUE
##  [40,]  TRUE  TRUE  TRUE  TRUE
##  [41,]  TRUE  TRUE  TRUE  TRUE
##  [42,]  TRUE  TRUE  TRUE  TRUE
##  [43,]  TRUE  TRUE  TRUE  TRUE
##  [44,]  TRUE  TRUE  TRUE  TRUE
##  [45,]  TRUE  TRUE  TRUE  TRUE
##  [46,]  TRUE  TRUE  TRUE  TRUE
##  [47,]  TRUE  TRUE  TRUE  TRUE
##  [48,]  TRUE  TRUE  TRUE  TRUE
##  [49,]  TRUE  TRUE  TRUE  TRUE
##  [50,]  TRUE  TRUE  TRUE  TRUE
##  [51,]  TRUE  TRUE  TRUE  TRUE
##  [52,]  TRUE  TRUE  TRUE  TRUE
##  [53,]  TRUE  TRUE  TRUE  TRUE
##  [54,]  TRUE  TRUE  TRUE  TRUE
##  [55,]  TRUE  TRUE  TRUE  TRUE
##  [56,]  TRUE  TRUE  TRUE  TRUE
##  [57,]  TRUE  TRUE  TRUE  TRUE
##  [58,]  TRUE  TRUE  TRUE  TRUE
##  [59,]  TRUE  TRUE  TRUE  TRUE
##  [60,]  TRUE  TRUE  TRUE  TRUE
##  [61,]  TRUE  TRUE  TRUE  TRUE
##  [62,]  TRUE  TRUE  TRUE  TRUE
##  [63,]  TRUE  TRUE  TRUE  TRUE
##  [64,]  TRUE  TRUE  TRUE  TRUE
##  [65,]  TRUE  TRUE  TRUE  TRUE
##  [66,]  TRUE  TRUE  TRUE  TRUE
##  [67,]  TRUE  TRUE  TRUE  TRUE
##  [68,]  TRUE  TRUE  TRUE  TRUE
##  [69,]  TRUE  TRUE  TRUE  TRUE
##  [70,]  TRUE  TRUE  TRUE  TRUE
##  [71,]  TRUE  TRUE  TRUE  TRUE
##  [72,]  TRUE  TRUE  TRUE  TRUE
##  [73,]  TRUE  TRUE  TRUE  TRUE
##  [74,]  TRUE  TRUE  TRUE  TRUE
##  [75,]  TRUE  TRUE  TRUE  TRUE
##  [76,]  TRUE  TRUE  TRUE  TRUE
##  [77,]  TRUE  TRUE  TRUE  TRUE
##  [78,]  TRUE  TRUE  TRUE  TRUE
##  [79,]  TRUE  TRUE  TRUE  TRUE
##  [80,]  TRUE  TRUE  TRUE  TRUE
##  [81,]  TRUE  TRUE  TRUE  TRUE
##  [82,]  TRUE  TRUE  TRUE  TRUE
##  [83,]  TRUE  TRUE  TRUE  TRUE
##  [84,]  TRUE  TRUE  TRUE  TRUE
##  [85,]  TRUE  TRUE  TRUE  TRUE
##  [86,]  TRUE  TRUE  TRUE  TRUE
##  [87,]  TRUE  TRUE  TRUE  TRUE
##  [88,]  TRUE  TRUE  TRUE  TRUE
##  [89,]  TRUE  TRUE  TRUE  TRUE
##  [90,]  TRUE  TRUE  TRUE  TRUE
##  [91,]  TRUE  TRUE  TRUE  TRUE
##  [92,]  TRUE  TRUE  TRUE  TRUE
##  [93,]  TRUE  TRUE  TRUE  TRUE
##  [94,]  TRUE  TRUE  TRUE  TRUE
##  [95,]  TRUE  TRUE  TRUE  TRUE
##  [96,]  TRUE  TRUE  TRUE  TRUE
##  [97,]  TRUE  TRUE  TRUE  TRUE
##  [98,]  TRUE  TRUE  TRUE  TRUE
##  [99,]  TRUE  TRUE  TRUE  TRUE
## [100,]  TRUE  TRUE  TRUE  TRUE
## [101,]  TRUE  TRUE  TRUE  TRUE
## [102,]  TRUE  TRUE  TRUE  TRUE
## [103,]  TRUE  TRUE  TRUE  TRUE
## [104,]  TRUE  TRUE  TRUE  TRUE
## [105,]  TRUE  TRUE  TRUE  TRUE
## [106,]  TRUE  TRUE  TRUE  TRUE
## [107,]  TRUE  TRUE  TRUE  TRUE
## [108,]  TRUE  TRUE  TRUE  TRUE
## [109,]  TRUE  TRUE  TRUE  TRUE
## [110,]  TRUE  TRUE  TRUE  TRUE
## [111,]  TRUE  TRUE  TRUE  TRUE
## [112,]  TRUE  TRUE  TRUE  TRUE
## [113,]  TRUE  TRUE  TRUE  TRUE
## [114,]  TRUE  TRUE  TRUE  TRUE
## [115,]  TRUE  TRUE  TRUE  TRUE
## [116,]  TRUE  TRUE  TRUE  TRUE
## [117,]  TRUE  TRUE  TRUE  TRUE
## [118,]  TRUE  TRUE  TRUE  TRUE
## [119,]  TRUE  TRUE  TRUE  TRUE
## [120,]  TRUE  TRUE  TRUE  TRUE
## [121,]  TRUE  TRUE  TRUE  TRUE
## [122,]  TRUE  TRUE  TRUE  TRUE
## [123,]  TRUE  TRUE  TRUE  TRUE
## [124,]  TRUE  TRUE  TRUE  TRUE
## [125,]  TRUE  TRUE  TRUE  TRUE
## [126,]  TRUE  TRUE  TRUE  TRUE
## [127,]  TRUE  TRUE  TRUE  TRUE
## [128,]  TRUE  TRUE  TRUE  TRUE
## [129,]  TRUE  TRUE  TRUE  TRUE
## [130,]  TRUE  TRUE  TRUE  TRUE
## [131,]  TRUE  TRUE  TRUE  TRUE
## [132,]  TRUE  TRUE  TRUE  TRUE
## [133,]  TRUE  TRUE  TRUE  TRUE
## [134,]  TRUE  TRUE  TRUE  TRUE
## [135,]  TRUE  TRUE  TRUE  TRUE
## [136,]  TRUE  TRUE  TRUE  TRUE
## [137,]  TRUE  TRUE  TRUE  TRUE
## [138,]  TRUE  TRUE  TRUE  TRUE
## [139,]  TRUE  TRUE  TRUE  TRUE
## [140,]  TRUE  TRUE  TRUE  TRUE
## [141,]  TRUE  TRUE  TRUE  TRUE
## [142,]  TRUE  TRUE  TRUE  TRUE
## [143,]  TRUE  TRUE  TRUE  TRUE
## [144,]  TRUE  TRUE  TRUE  TRUE
## [145,]  TRUE  TRUE  TRUE  TRUE
## [146,]  TRUE  TRUE  TRUE  TRUE
## [147,]  TRUE  TRUE  TRUE  TRUE
## [148,]  TRUE  TRUE  TRUE  TRUE
## [149,]  TRUE  TRUE  TRUE  TRUE
## [150,]  TRUE  TRUE  TRUE  TRUE
## [151,]  TRUE  TRUE  TRUE  TRUE
## [152,]  TRUE  TRUE  TRUE  TRUE
## [153,]  TRUE  TRUE  TRUE  TRUE
## [154,]  TRUE  TRUE  TRUE  TRUE
## [155,]  TRUE  TRUE  TRUE  TRUE
## [156,]  TRUE  TRUE  TRUE  TRUE
## [157,]  TRUE  TRUE  TRUE  TRUE
## [158,]  TRUE  TRUE  TRUE  TRUE
## [159,]  TRUE  TRUE  TRUE  TRUE
## [160,]  TRUE  TRUE  TRUE  TRUE
## [161,]  TRUE  TRUE  TRUE  TRUE
## [162,]  TRUE  TRUE  TRUE  TRUE
## [163,]  TRUE  TRUE  TRUE  TRUE
## [164,]  TRUE  TRUE  TRUE  TRUE
## [165,]  TRUE  TRUE  TRUE  TRUE
## [166,]  TRUE  TRUE  TRUE  TRUE
## [167,]  TRUE  TRUE  TRUE  TRUE
## [168,]  TRUE  TRUE  TRUE  TRUE
## [169,]  TRUE  TRUE  TRUE  TRUE
## [170,]  TRUE  TRUE  TRUE  TRUE
## [171,]  TRUE  TRUE  TRUE  TRUE
## [172,]  TRUE  TRUE  TRUE  TRUE
## [173,]  TRUE  TRUE  TRUE  TRUE
## [174,]  TRUE  TRUE  TRUE  TRUE
## [175,]  TRUE  TRUE  TRUE  TRUE
## [176,]  TRUE  TRUE  TRUE  TRUE
## [177,]  TRUE  TRUE  TRUE  TRUE
## [178,]  TRUE  TRUE  TRUE  TRUE
## [179,]  TRUE  TRUE  TRUE  TRUE
## [180,]  TRUE  TRUE  TRUE  TRUE
## [181,]  TRUE  TRUE  TRUE  TRUE
## [182,]  TRUE  TRUE  TRUE  TRUE
## [183,]  TRUE  TRUE  TRUE  TRUE
## [184,]  TRUE  TRUE  TRUE  TRUE
## [185,]  TRUE  TRUE  TRUE  TRUE
## [186,]  TRUE  TRUE  TRUE  TRUE
## [187,]  TRUE  TRUE  TRUE  TRUE
## [188,]  TRUE  TRUE  TRUE  TRUE
## [189,]  TRUE  TRUE  TRUE  TRUE
## [190,]  TRUE  TRUE  TRUE  TRUE
## [191,]  TRUE  TRUE  TRUE  TRUE
## [192,]  TRUE  TRUE  TRUE  TRUE
## [193,]  TRUE  TRUE  TRUE  TRUE
## [194,]  TRUE  TRUE  TRUE  TRUE
## [195,]  TRUE  TRUE  TRUE  TRUE
## [196,]  TRUE  TRUE  TRUE  TRUE
## [197,]  TRUE  TRUE  TRUE  TRUE
## [198,]  TRUE  TRUE  TRUE  TRUE
## [199,]  TRUE  TRUE  TRUE  TRUE
## [200,]  TRUE  TRUE  TRUE  TRUE
## [201,]  TRUE  TRUE  TRUE  TRUE
## [202,]  TRUE  TRUE  TRUE  TRUE
## [203,]  TRUE  TRUE  TRUE  TRUE
## [204,]  TRUE  TRUE  TRUE  TRUE
## [205,]  TRUE  TRUE  TRUE  TRUE
## [206,]  TRUE  TRUE  TRUE  TRUE
## [207,]  TRUE  TRUE  TRUE  TRUE
## [208,]  TRUE  TRUE  TRUE  TRUE
## [209,]  TRUE  TRUE  TRUE  TRUE
## [210,]  TRUE  TRUE  TRUE  TRUE
## [211,]  TRUE  TRUE  TRUE  TRUE
## [212,]  TRUE  TRUE  TRUE  TRUE
## [213,]  TRUE  TRUE  TRUE  TRUE
## [214,]  TRUE  TRUE  TRUE  TRUE
## [215,]  TRUE  TRUE  TRUE  TRUE
## [216,]  TRUE  TRUE  TRUE  TRUE
## [217,]  TRUE  TRUE  TRUE  TRUE
## [218,]  TRUE  TRUE  TRUE  TRUE
## [219,]  TRUE  TRUE  TRUE  TRUE
## [220,]  TRUE  TRUE  TRUE  TRUE
## [221,]  TRUE  TRUE  TRUE  TRUE
## [222,]  TRUE  TRUE  TRUE  TRUE
## [223,]  TRUE  TRUE  TRUE  TRUE
## [224,]  TRUE  TRUE  TRUE  TRUE
## [225,]  TRUE  TRUE  TRUE  TRUE
## [226,]  TRUE  TRUE  TRUE  TRUE
## [227,]  TRUE  TRUE  TRUE  TRUE
## [228,]  TRUE  TRUE  TRUE  TRUE
## [229,]  TRUE  TRUE  TRUE  TRUE
## [230,]  TRUE  TRUE  TRUE  TRUE
## [231,]  TRUE  TRUE  TRUE  TRUE
## [232,]  TRUE  TRUE  TRUE  TRUE
## [233,]  TRUE  TRUE  TRUE  TRUE
## [234,]  TRUE  TRUE  TRUE  TRUE
```

We see that, in the vast majority of cases, the different elements of the dataset seem to be correlated. However, due to the nature of the lower triangular matrix, it is very difficult to determine whether the relationship here is positive or negative.