Car Fuel Efficiency Cleaning

Harrison Bailye

08/12/2021

Contents

Load in the data

Data from my Subaru Impreza was collected using an OBD reader and was exported into a csv file. The data is exported in monthly blocks so each file needs to be imported separately and then joined to make one big data table.

```
car_10 <- read_csv("driving_2021_10.csv")
```

```
## Rows: 37 Columns: 23
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (14): Driving Start Time, Driving Finish Time, Driving Time, Address, Di...
## dbl  (9): Fuel Consumed, Fuel Cost, Rapid Acc. Hard Count, Rapid Acc. Normal...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
car_11 <- read_csv("driving_2021_11.csv")
```

```
## Rows: 57 Columns: 23
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (14): Driving Start Time, Driving Finish Time, Driving Time, Address, Di...
## dbl  (9): Fuel Consumed, Fuel Cost, Rapid Acc. Hard Count, Rapid Acc. Normal...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
car_12 <- read_csv("driving_2021_12.csv")
```

```
## Rows: 84 Columns: 23
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (14): Driving Start Time, Driving Finish Time, Driving Time, Address, Di...
## dbl  (9): Fuel Consumed, Fuel Cost, Rapid Acc. Hard Count, Rapid Acc. Normal...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Join the Data

We will join all the data together to create one big data table.

```
car <- full_join(car_10, car_11)
```

```
## Joining, by = c("Driving Start Time", "Driving Finish Time", "Driving Time", "Address", "Distance",
```

```
car <- full_join(car, car_12)
```

```
## Joining, by = c("Driving Start Time", "Driving Finish Time", "Driving Time", "Address", "Distance",
```

EDA

Before a model can be constructed, it is important to perform explanatory data analysis (EDA) to ensure that the data is clean and in the correct form and also to investigate relationships between predictors and the response variable.

EDA 01: View the data

To begin with, we will view the data to see what needs to be fixed before working with the data.

```
head(car)
```

```
## # A tibble: 6 x 23
##   `Driving Start T~ `Driving Finish~ `Driving Time` Address Distance `Max Speed`
##   <chr>             <chr>            <chr>          <chr>   <chr>    <chr>
## 1 2021.10.19 16:52~ 2021.10.19 16:5~ 6m 48s         22 Win~ 0.75km   42.0km/h
## 2 2021.10.20 19:07~ 2021.10.20 19:2~ 14m 34s        71 Dev~ 8.95km   62.0km/h
## 3 2021.10.20 21:41~ 2021.10.20 21:5~ 12m 8s         22 Win~ 8.75km   64.0km/h
## 4 2021.10.21 19:32~ 2021.10.21 19:4~ 8m 2s          1-9 Be~ 4.54km   60.0km/h
## 5 2021.10.22 07:45~ 2021.10.22 08:2~ 43m 5s         1-15 H~ 17.18km  64.0km/h
## 6 2021.10.22 08:45~ 2021.10.22 09:1~ 27m 43s        119 Be~ 16.08km  66.0km/h
## # ... with 17 more variables: Avr. Speed <chr>, Max RPM <chr>, Avr. RPM <chr>,
## #   Max Coolant Temp <chr>, Max Engine Oil Temp <chr>, Fuel effciency <chr>,
## #   Fuel Consumed <dbl>, Fuel Cost <dbl>, Rapid Acc. Hard Count <dbl>,
## #   Rapid Acc. Normal Count <dbl>, Rapid Decel. Hard Count <dbl>,
## #   Rapid Decel. Normal Count <dbl>, Speeding Hard count <dbl>,
## #   Speeding Normal count <dbl>, Idling time <dbl>, Safe Driving Score <chr>,
## #   Eco Driving Score <chr>
```

We can omit the missing data points as there are not many of them so omitting them won't have much of an impact on the results.

```
car <- car %>%
  na.omit()
```

EDA 02: Remove predictors

We saw that there were 22 predictors in the data set, however, not all of them are useful for the purpose of the task, so we will omit these predictors from the data set.

```
car <- car %>%
  select(-c(Address, `Driving Start Time`, `Driving Finish Time`, `Max Engine Oil Temp`,
            `Fuel Cost`, `Speeding Normal count`, `Speeding Hard count`,
            `Rapid Acc. Hard Count`, `Rapid Acc. Normal Count`, `Driving Time`))
head(car)
```

```
## # A tibble: 6 x 13
##   Distance `Max Speed` `Avr. Speed` `Max RPM`  `Avr. RPM` `Max Coolant Temp`
##   <chr>    <chr>       <chr>        <chr>      <chr>      <chr>
## 1 0.75km   42.0km/h    6.66km/h     2346.75rpm 989.55rpm  92.0°C
## 2 8.95km   62.0km/h    37.41km/h    3037.5rpm  1549.8rpm  97.0°C
```

```
## 3 8.75km   64.0km/h    43.28km/h    2684.25rpm 1543.91rpm 97.0°C
## 4 4.54km   60.0km/h    33.93km/h    2775.0rpm  1468.14rpm 96.0°C
## 5 17.18km  64.0km/h    23.94km/h    2693.75rpm 1228.58rpm 98.0°C
## 6 16.08km  66.0km/h    34.82km/h    2634.25rpm 1347.62rpm 98.0°C
## # ... with 7 more variables: Fuel effciency <chr>, Fuel Consumed <dbl>,
## #   Rapid Decel. Hard Count <dbl>, Rapid Decel. Normal Count <dbl>,
## #   Idling time <dbl>, Safe Driving Score <chr>, Eco Driving Score <chr>
```

EDA 03: Clean columns

We will now clean the columns, converting them into the right form and removing the units of measurement for each data point.

```
# Remove units of measurement
car$Distance <- str_remove_all(car$Distance, "km")
car$`Avr. Speed` <- str_remove_all(car$`Avr. Speed`, "km/h")
car$`Max Speed` <- str_remove_all(car$`Max Speed`, "km/h")
car$`Avr. RPM` <- str_remove_all(car$`Avr. RPM`, "rpm")
car$`Max RPM` <- str_remove_all(car$`Max RPM`, "rpm")
car$`Max Coolant Temp` <- str_remove_all(car$`Max Coolant Temp`, "°C")
car$`Fuel effciency` <- str_remove_all(car$`Fuel effciency`, "L/100km")

# Convert to numeric
car$Distance <-  as.numeric(car$Distance)
car$`Avr. Speed` <- as.numeric(car$`Avr. Speed`)
car$`Max Speed` <- as.numeric(car$`Max Speed`)
car$`Avr. RPM` <- as.numeric(car$`Avr. RPM`)
car$`Max RPM` <- as.numeric(car$`Max RPM`)
car$`Max Coolant Temp` <- as.numeric(car$`Max Coolant Temp`)
car$`Fuel effciency` <- as.numeric(car$`Fuel effciency`)
car$`Safe Driving Score` <- as.numeric(car$`Safe Driving Score`)
```

```
## Warning: NAs introduced by coercion
```

```
car$`Eco Driving Score` <- as.numeric(car$`Eco Driving Score`)
```

```
## Warning: NAs introduced by coercion
```

Viewing the Clean Data

We can see the final clean data below.

```
head(car)
```

```
## # A tibble: 6 x 13
##   Distance `Max Speed` `Avr. Speed` `Max RPM` `Avr. RPM` `Max Coolant Temp`
##      <dbl>       <dbl>        <dbl>     <dbl>      <dbl>              <dbl>
## 1     0.75          42         6.66     2347.       990.                 92
## 2     8.95          62        37.4      3038.      1550.                 97
## 3     8.75          64        43.3      2684.      1544.                 97
## 4     4.54          60        33.9      2775       1468.                 96
```

```
## 5    17.2           64        23.9      2694.      1229.                 98
## 6    16.1           66        34.8      2634.      1348.                 98
## # ... with 7 more variables: Fuel effciency <dbl>, Fuel Consumed <dbl>,
## #   Rapid Decel. Hard Count <dbl>, Rapid Decel. Normal Count <dbl>,
## #   Idling time <dbl>, Safe Driving Score <dbl>, Eco Driving Score <dbl>
```

Saving the Data

```
write.csv(car, file = "car_data.csv")
```