Car Fuel Consumption on Kaggle

Joshua Manalo

October 31, 2018

Contents

1. Introduction

The following data came from Kaggle. The owner of the data obtained the observations from the car's display after each ride while regularly changing the gas type.

The main question of the owner is to know which of the 2 gases are cheaper? E10(SP95) or Sp98? E10 (SP95) is "super gas" with 10% Ethanol, and is sold for 1.38 Euros/Litre While SP98 is sold for 1.46 Euros/Litre

The data analysis will be done in Rstudio and Rmarkdown.

2. Data Structure

In the data we have the following variables:

Variable	Description
Distance	distance travelled in kilometres (kms)
Consume	Litres per 100 kms (L/100km)
Speed	Average speed in kilometres per hour (km/h)
Temp_inside	The temperature inside the car (c)
Temp_outside	The temperature outside the car (c)
Gas Type	E10 and SP98
AC	1 if on; 0 if off
Rain	1 if on; 0 if off
Sun	1 if sunny enough; 0 if not sunny/hot enough

There are **388 observations**. Therefore there was 388 trips made by the owner at the time of this data analysis.

3.Preparations

I decided to focus on **Consume** as my target variable. To start my analysis, I checked which of the variables are highly significant to Consumption

First I **divided the data into 2 gas types (E10 and SP98)**. By fitting a regression with Consumption as my dependent variable.

```
e10.fit <- lm(consume ~ speed+distance+temp_inside+temp_outside+AC+rain+sun,data=e10.data)
round(summary(e10.fit)$coeff,3)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.063      2.662   1.526    0.129
speed          -0.014      0.006  -2.213    0.028
distance        0.001      0.004   0.160    0.873
temp_inside     0.084      0.122   0.686    0.494
temp_outside   -0.041      0.012  -3.516    0.001
AC              0.010      0.363   0.027    0.978
rain            0.615      0.236   2.608    0.010
sun            -0.192      0.284  -0.675    0.501
```

```
sp98.fit <- lm(consume ~ speed+distance+temp_inside+temp_outside+AC+rain+sun,data=sp98.data)
round(summary(sp98.fit)$coeff,3)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.394      1.369   6.132    0.000
speed          -0.021      0.006  -3.332    0.001
distance        0.002      0.003   0.638    0.524
temp_inside    -0.105      0.064  -1.632    0.104
temp_outside   -0.042      0.011  -3.723    0.000
AC              0.575      0.248   2.317    0.021
rain            0.461      0.212   2.173    0.031
sun            -0.096      0.264  -0.363    0.717
```

From my regression models, it shows that the variables that are significant to Consumes are **Speed**, **Temperature Outside** and **Rain**.

3.1. Temperature Outside Variable

When comparing temperature outside recorded between the 2 gas types, I noticed the following:

```
E10 Temperature Outside Observations Summary
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -5.00    6.00    9.00   10.12   14.25   27.00
```

```
SP98 Temperature Outside Observations Summary
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -3.00    7.00   11.00   12.23   17.00   31.00
```

```
  t.test.statistic t.test.p.value
t        -2.954438    0.003324326
```

Since **p-value $<$ 0.05** We reject the null and confirm that **there is a significant difference between the E10 and SP98 temperature outside means**. This tells us that both gas types have not been used in similar temperatures. Therefore I standardize the variable by determining the range of temperature outside we will observe.

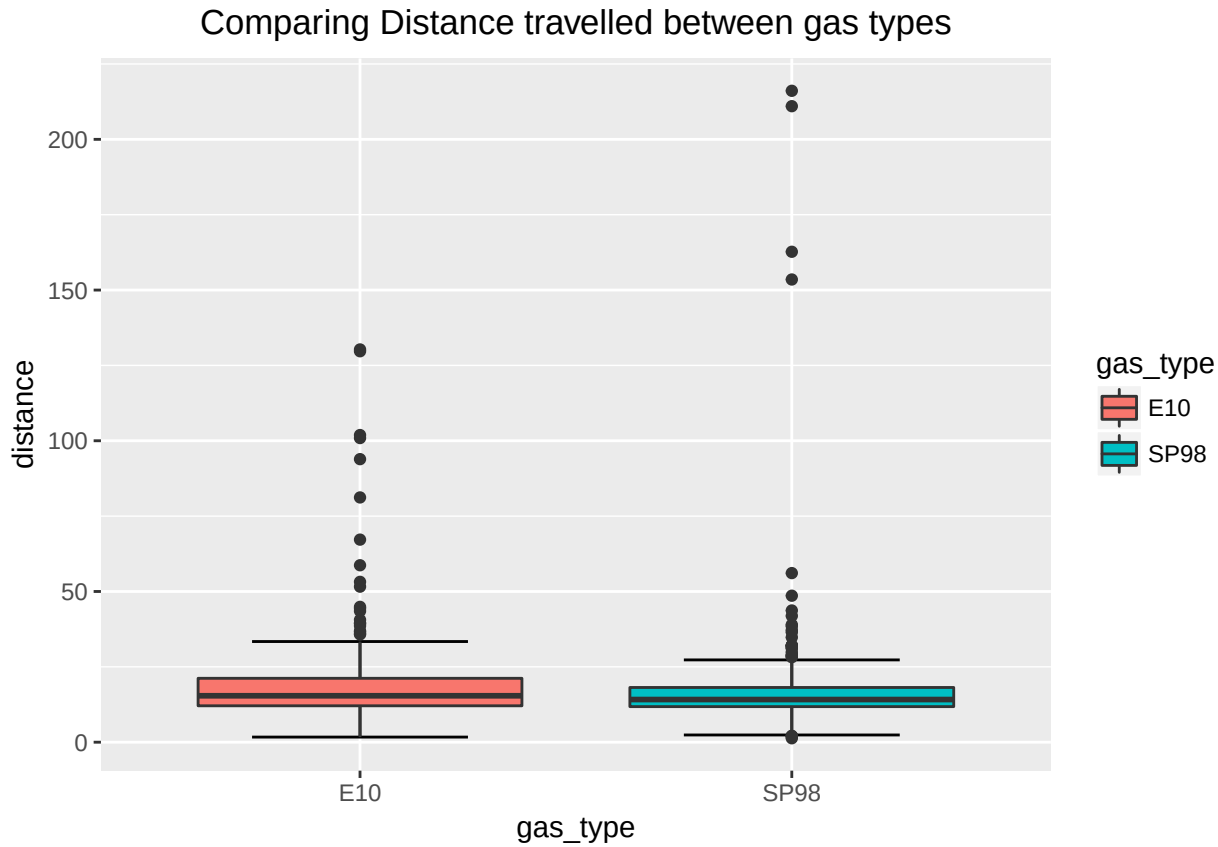
3.2. Distance Variable

```
E10 Distance Observations Summary
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.70   12.07   15.40   21.10   21.20  130.30
```

```
SP98 Distance Observations Summary
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.30   11.80   14.15   18.64   18.15  216.10
```



We see that there is quite a **few outliers from the E10 and SP98 distance observations**. This can influence the consume variable (Since consume is L/100kms) so we standardize the variable by determining the range of distance we will observe.

3.3. Rain Variable

```
E10 Rain count
```

```
  0   1
144  16
```

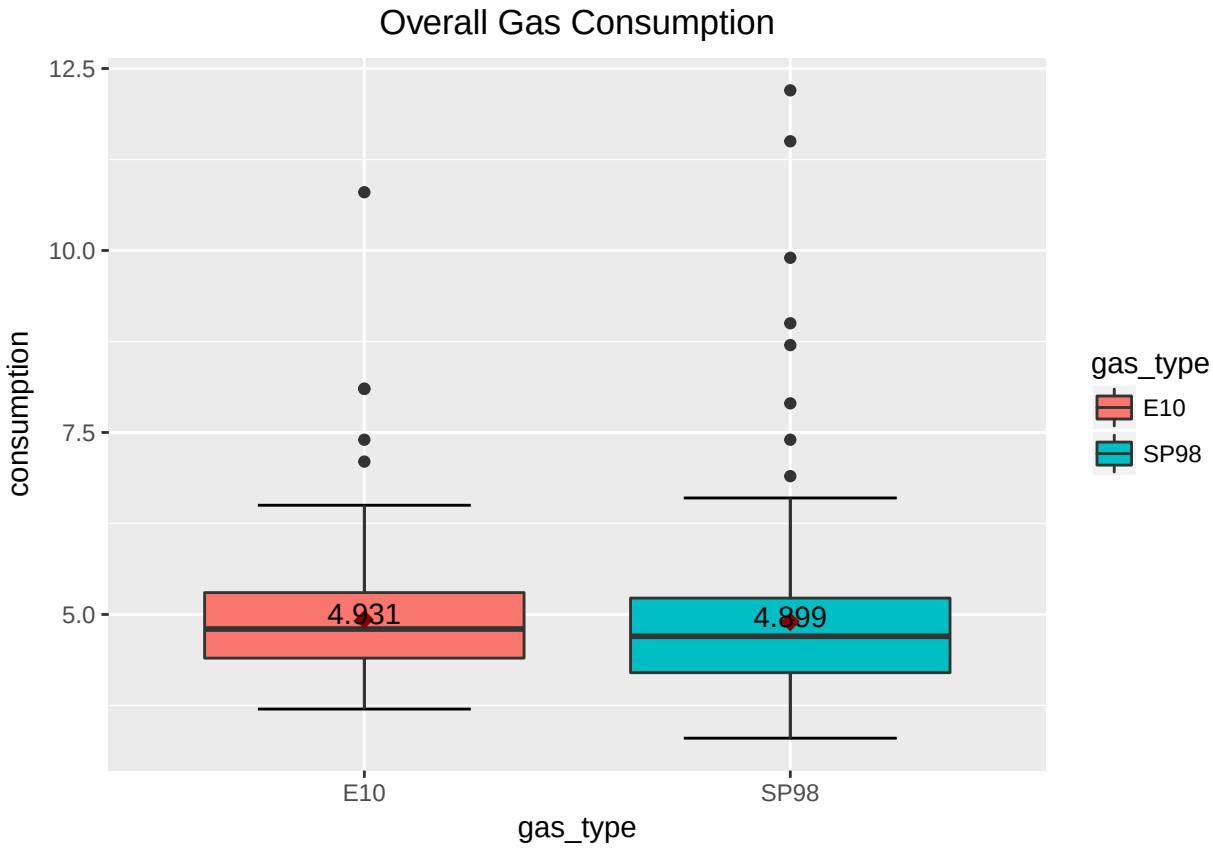
```
SP98 Rain count
```

```
  0   1
196  32
```

We notice that **not a lot of data have been recorded with rain**. This is quite unfortunate since rain is a factor that has an impact on the consume variable

4. Simple Visualizations

4.1. Consumption

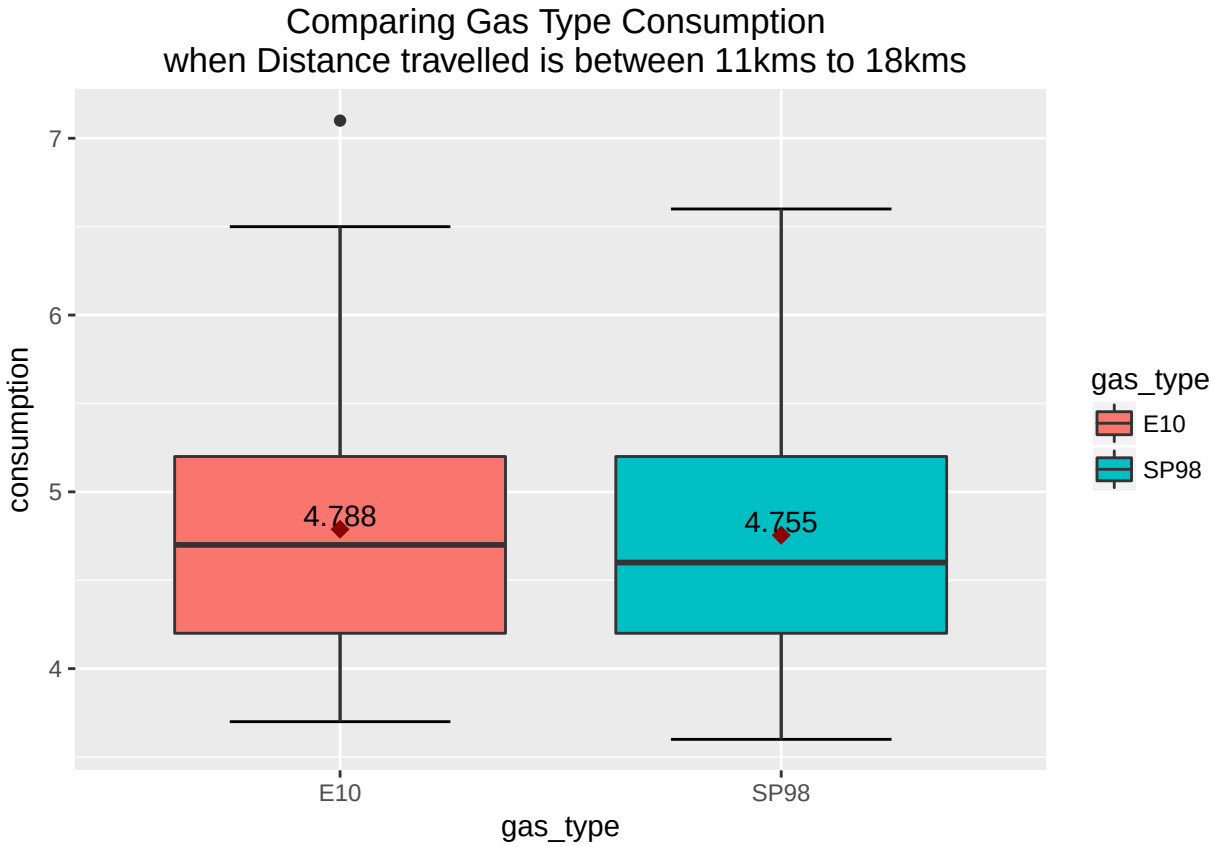


There is slightly higher average consumption for **E10 (4.931L/100km)** than **SP98 (4.899L/100km)** when we include all observations.

4.2. Gas Consumption with distance

Earlier we saw outliers for distance between gas types. This can influence consumption so we standardize the variable by determining the range of distance we will observe.

I decide to **restrict the range between 11kms to 18kms**.

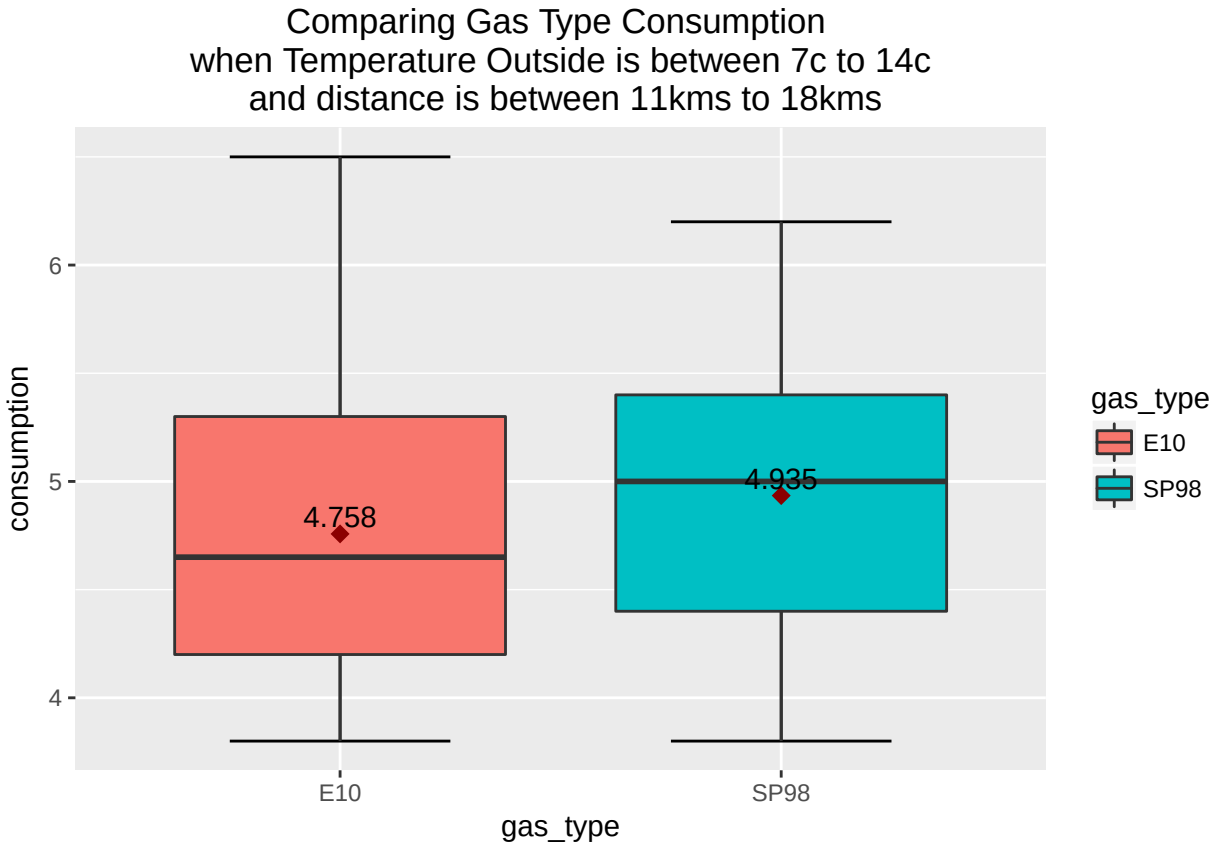


There is slightly higher average consumption for **E10 (4.788L/100km)** than **SP98 (4.755L/100km)** when we only use data observations where distance travelled is between 11kms to 18kms

4.3. Gas Consumption with distance and temperature outside

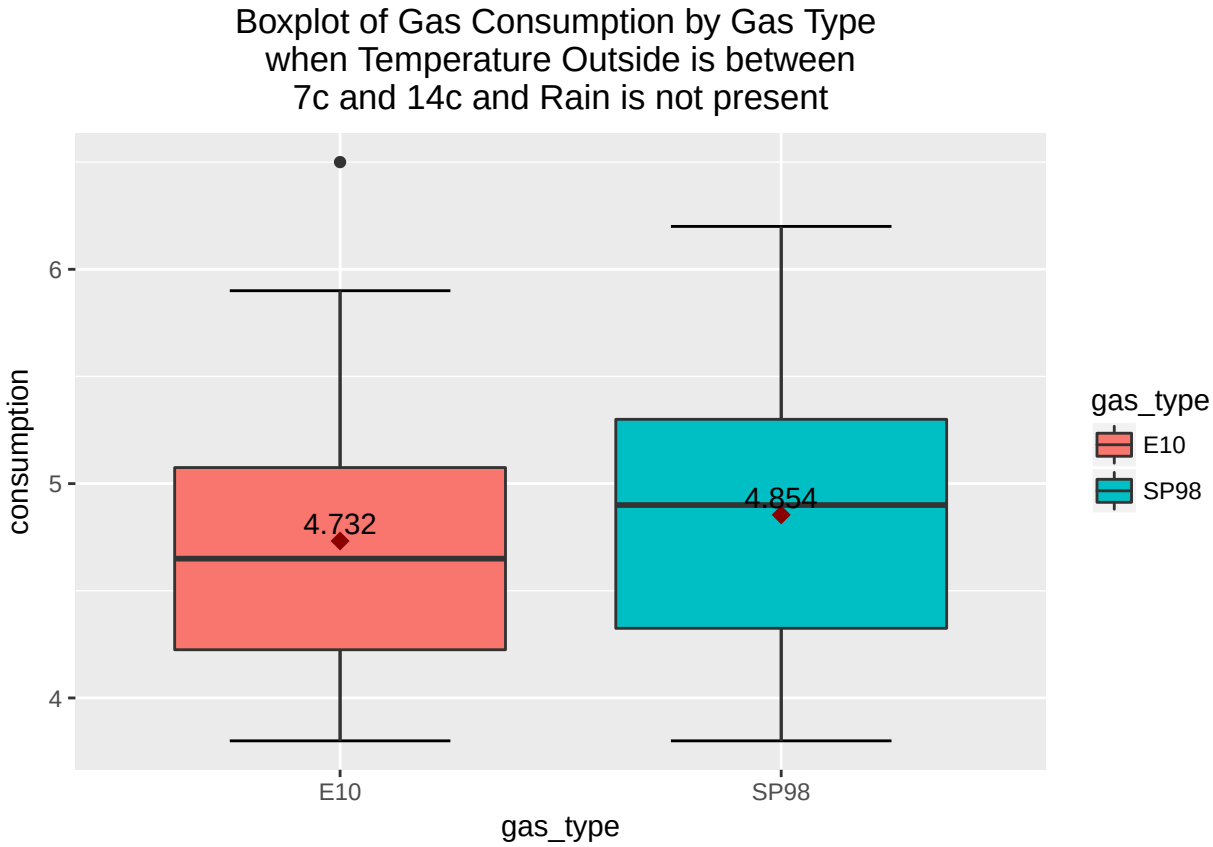
We check what happens when temp outside is standardized. We fix the range because earlier saw the means of temperature outside between the fuel consumption were not equal.

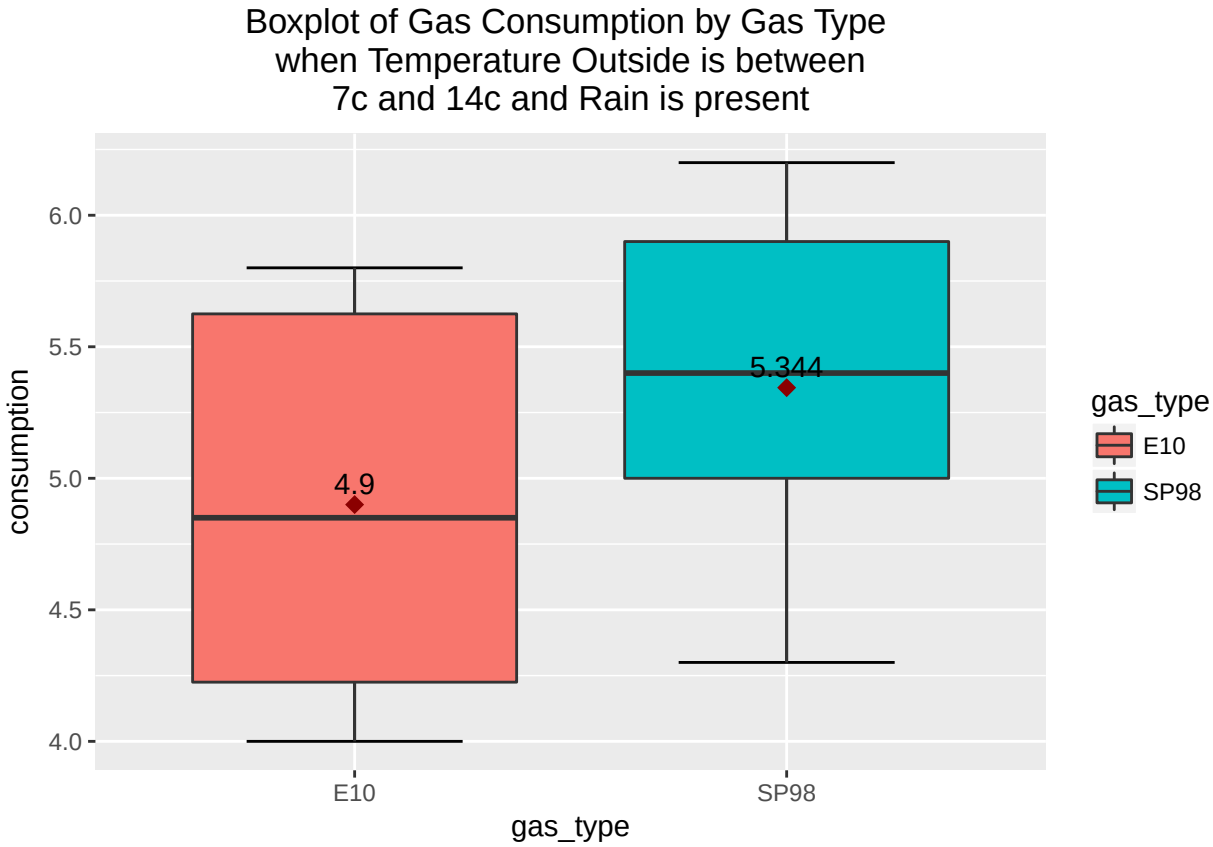
I decide to **restrict the range between 7c to 14c**.



There is slightly higher average consumption for **SP98 (4.935L/100km)** than **E10 (4.758L/100km)** when we only use data between 7c to 14c and distance travelled is between 11kms to 18kms

4.4. Gas consumption with distance, temperature outside, and rain





There is slightly higher average consumption for **SP98 (4.854L/100km)** than **E10 (4.732L/100km)** when we only use data between 7c to 14c, distance travelled is between 11kms to 18kms, and not raining.

There is slightly higher average consumption for **SP98 (5.344L/100km)** than **E10 (4.9L/100km)** when we only use data between 7c to 14c, distance travelled is between 11kms to 18kms, and is raining.

*Please note the higher variance in the boxplot models as there are not many observations were recorded for when rain was present.

5. Conclusion

The first 2 graphs showed the gas type SP98 on average had less consumption than the E10 gas type. However when we take into account the differences in temperature outside and distance outliers by fixing its range, we see that E10 gas type consumes slightly less than SP98. This shows on our final 3 graphs.

Therefore we can safely conclude that E10 (SP95) is a much cheaper alternative gas for your car.

6. Future Project

Some departing ideas to build on, we can compare observations between gas types when temperature outside is in a much colder season (i.e between -20 to 0).
We can also analyze in depth our regression models and see which gas types are more affected by outside variables such as weather.