Statistical Tests with R: T-tests, ANOVA, and Correlation

Mohan Khanal

2025-08-05

Table of contents

1 Statistical Tests with R (T-tests, ANOVA, and correlation)

1.1 Load required libraries

```r
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag
```

```
The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(readr)
library(ggplot2)
```

```
## Import dataset ----
car_data <- read.csv("data/car data.csv")
#car_data <- read_csv("D:/DemoPracticedata/car data.csv")

# Preview data
head(car_data)
```

```
       Car_Name Year Selling_Price Present_Price Kms_Driven Fuel_Type
1          ritz 2014          3.35          5.59      27000    Petrol
2           sx4 2013          4.75          9.54      43000    Diesel
3          ciaz 2017          7.25          9.85       6900    Petrol
4       wagon r 2011          2.85          4.15       5200    Petrol
5         swift 2014          4.60          6.87      42450    Diesel
6 vitara brezza 2018          9.25          9.83       2071    Diesel
  Seller_Type Transmission Owner
1      Dealer       Manual     0
2      Dealer       Manual     0
3      Dealer       Manual     0
4      Dealer       Manual     0
5      Dealer       Manual     0
6      Dealer       Manual     0
```

1.2 Part 1: Exploratory Data Analysis

1.2.1 Cross-Tabulation

Cross-tabulation summarizes the relationship between two categorical variables, showing counts and percentages. Example: How do Fuel_Type and Transmission relate?

```
#row percentage
crosstab_rw <- modelsummary::datasummary_crosstab(
  Fuel_Type ~ Transmission,
  statistic = ~N +1 +Percent("row"),
  data = car_data
```

Fuel_Type		Automatic	Manual	All
CNG	N	0	2	2
	% row	0.0	100.0	100.0
Diesel	N	12	48	60
	% row	20.0	80.0	100.0
Petrol	N	28	211	239
	% row	11.7	88.3	100.0

Fuel_Type		Automatic	Manual	All
CNG	N	0	2	2
	% col	0.0	0.8	0.7
Diesel	N	12	48	60
	% col	30.0	18.4	19.9
Petrol	N	28	211	239
	% col	70.0	80.8	79.4

```
)
crosstab_rw
```

```
#column percentage
crosstab_cl <- modelsummary::datasummary_crosstab(
  Fuel_Type ~ Transmission,
  statistic = ~ N + 1+Percent("col"),
  data = car_data
)
crosstab_cl
```

1.2.2 Correlation

Correlation measures the strength and direction of the relationship between two numeric variables.

Example: Is there a relationship between Selling_Price and Kms_Driven?

```r
correlation <- cor.test(car_data$Selling_Price, car_data$Kms_Driven)
correlation
```

```
    Pearson's product-moment correlation

data:  car_data$Selling_Price and car_data$Kms_Driven
t = 0.50491, df = 299, p-value = 0.614
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.08414286  0.14177160
sample estimates:
       cor
0.02918709
```

Interpretation: Correlation coefficient (r): Positive (both increase together), negative (one increases, other decreases), or near zero (no linear relationship). p-value < 0.05 indicates a statistically significant relationship.

1.3 Part 2: T-tests

T-tests compare the means of two groups to determine if there is a statistically significant difference. Example: Is the average Selling_Price different between manual and automatic cars?

```r
t_test_result <- t.test(Selling_Price ~ Transmission, data = car_data)
t_test_result
```

```
    Welch Two Sample t-test

data:  Selling_Price by Transmission
t = 3.9055, df = 41.248, p-value = 0.0003417
alternative hypothesis: true difference in means between group Automatic and group Manual is
95 percent confidence interval:
 2.650698 8.325318
sample estimates:
mean in group Automatic    mean in group Manual
               9.420000                3.931992
```

Interpretation: p-value < 0.05: Significant difference in means. Confidence interval: Shows the range of the true difference in means. p-value > 0.05: No strong evidence of a difference.

##Part 3: ANOVA ANOVA compares means across three or more groups. Example: Does average Selling_Price differ among Fuel_Type categories?

```
anova_result <- aov(Selling_Price ~ Transmission, data = car_data)
summary(anova_result)
```

```
              Df Sum Sq Mean Sq F value  Pr(>F)
Transmission   1   1045  1044.6   46.58 4.9e-11 ***
Residuals    299   6706    22.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.3.1 Post-hoc Test

```
TukeyHSD(anova_result)
```

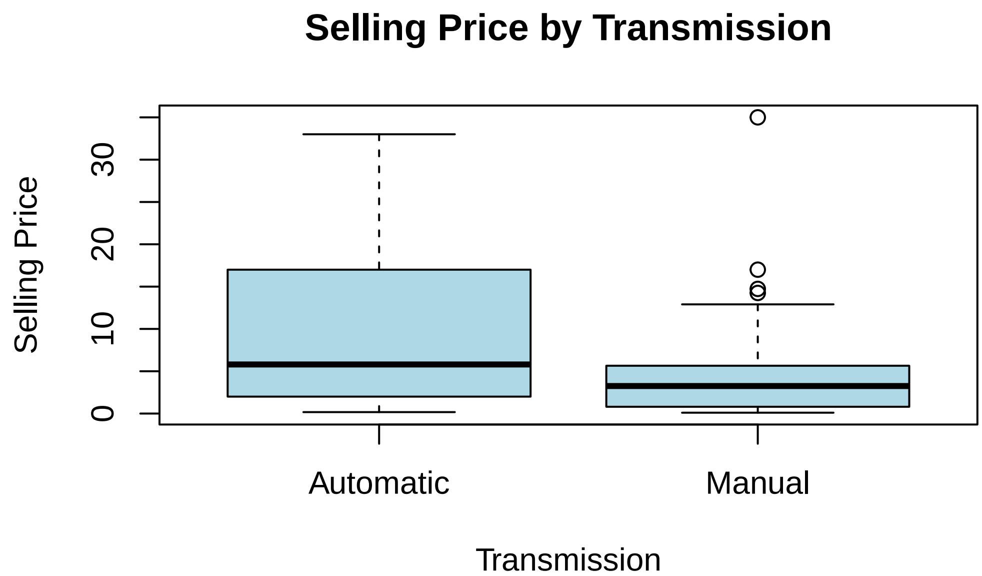
```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Selling_Price ~ Transmission, data = car_data)

$Transmission
                      diff       lwr       upr p adj
Manual-Automatic -5.488008 -7.070473 -3.905542     0
```

Interpretation:

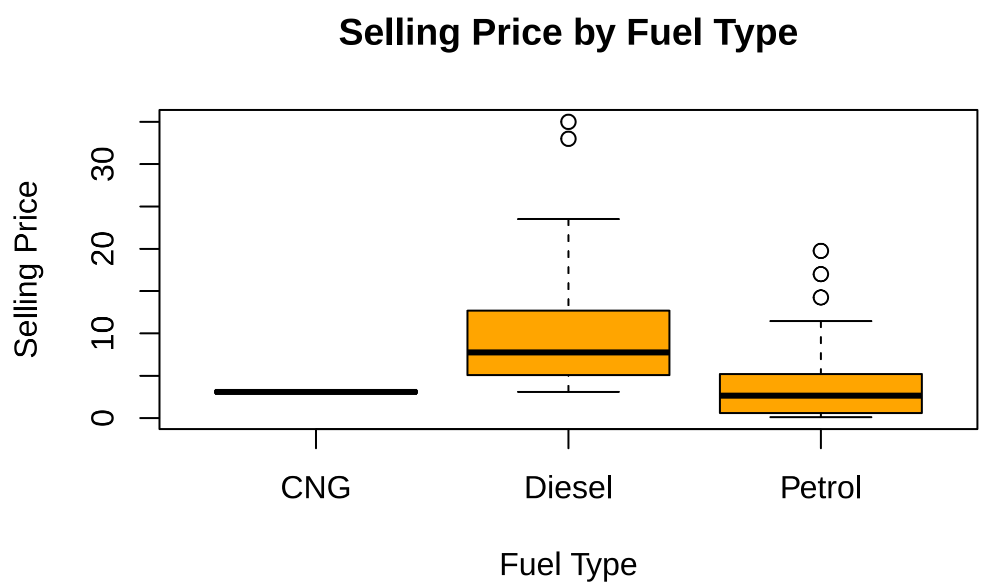
p-value < 0.05: At least one group mean differs. Tukey's test: Identifies which specific groups differ.

1.4 Part 4: Visualizations

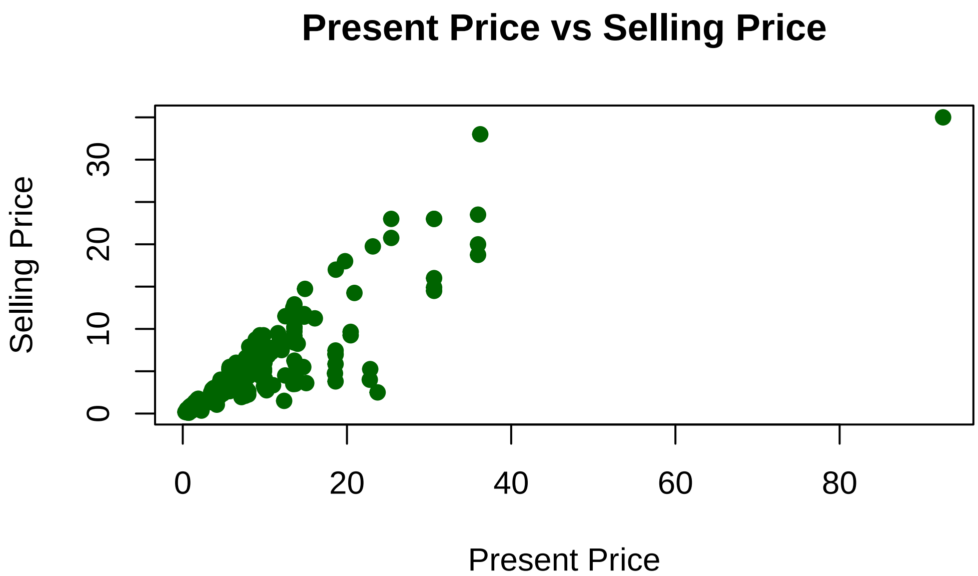
```
# Boxplot for T-test variable
boxplot(Selling_Price ~ Transmission, data = car_data,
        main = "Selling Price by Transmission",
        xlab = "Transmission", ylab = "Selling Price",
        col = "lightblue")
```




```
# Boxplot for ANOVA variable
boxplot(Selling_Price ~ Fuel_Type, data = car_data,
        main = "Selling Price by Fuel Type",
        xlab = "Fuel Type", ylab = "Selling Price",
        col = "orange")
```




```r
# Scatter plot for correlation
plot(car_data$Present_Price, car_data$Selling_Price,
     main = "Present Price vs Selling Price",
     xlab = "Present Price", ylab = "Selling Price",
     pch = 19, col = "darkgreen")
```


Present Price vs Selling Price
Selling Price
0
10
20
30
0
20
40
60
80
Present Price