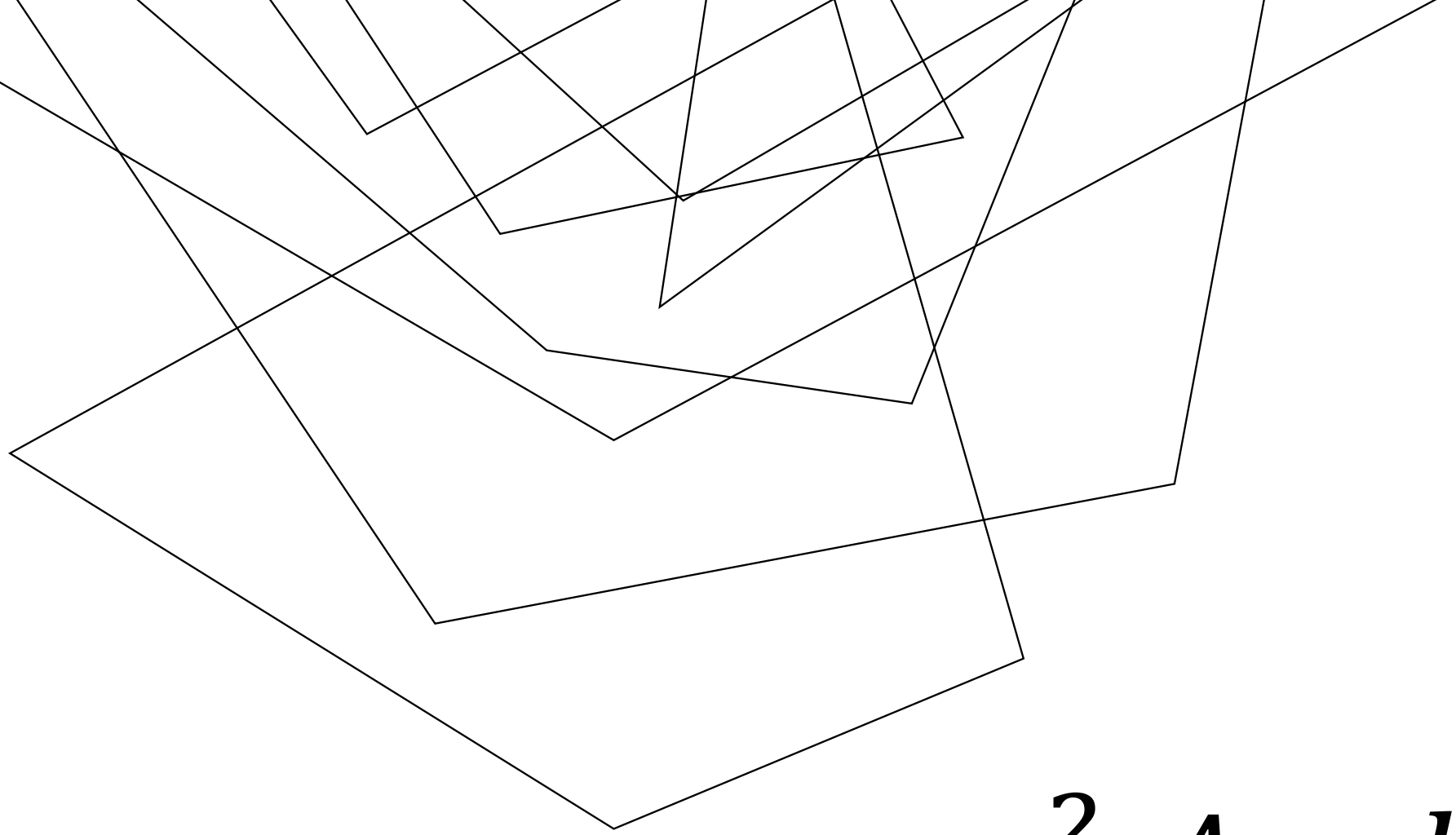
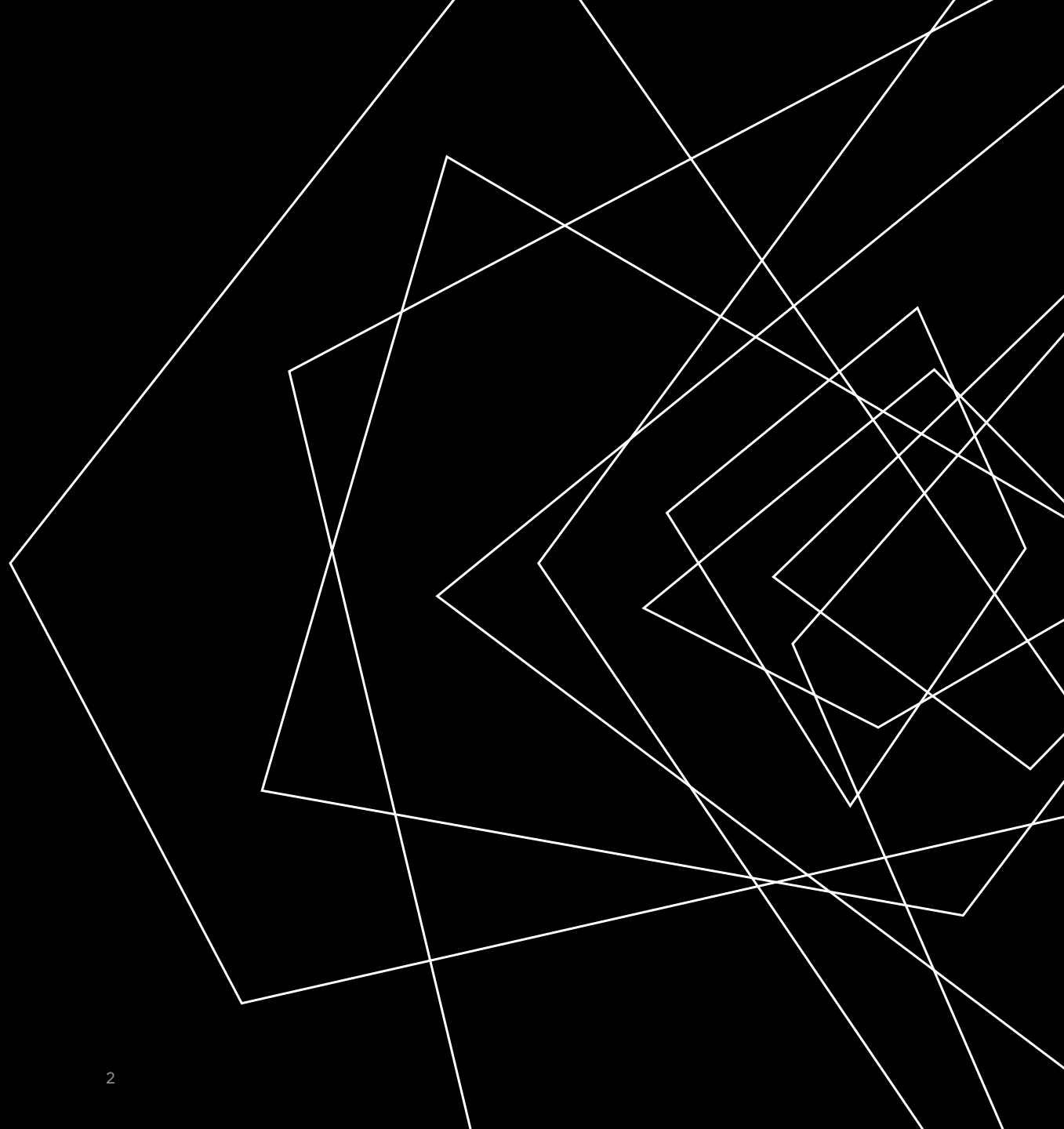


χ^2 Analysis

Anjali Savlani

AGENDA

- Introduction
- Primary Goal
- Analysis Description
- Insights from Analysis & Explanation
- Recommendation of two other Combinations
- Conclusion
- R Code



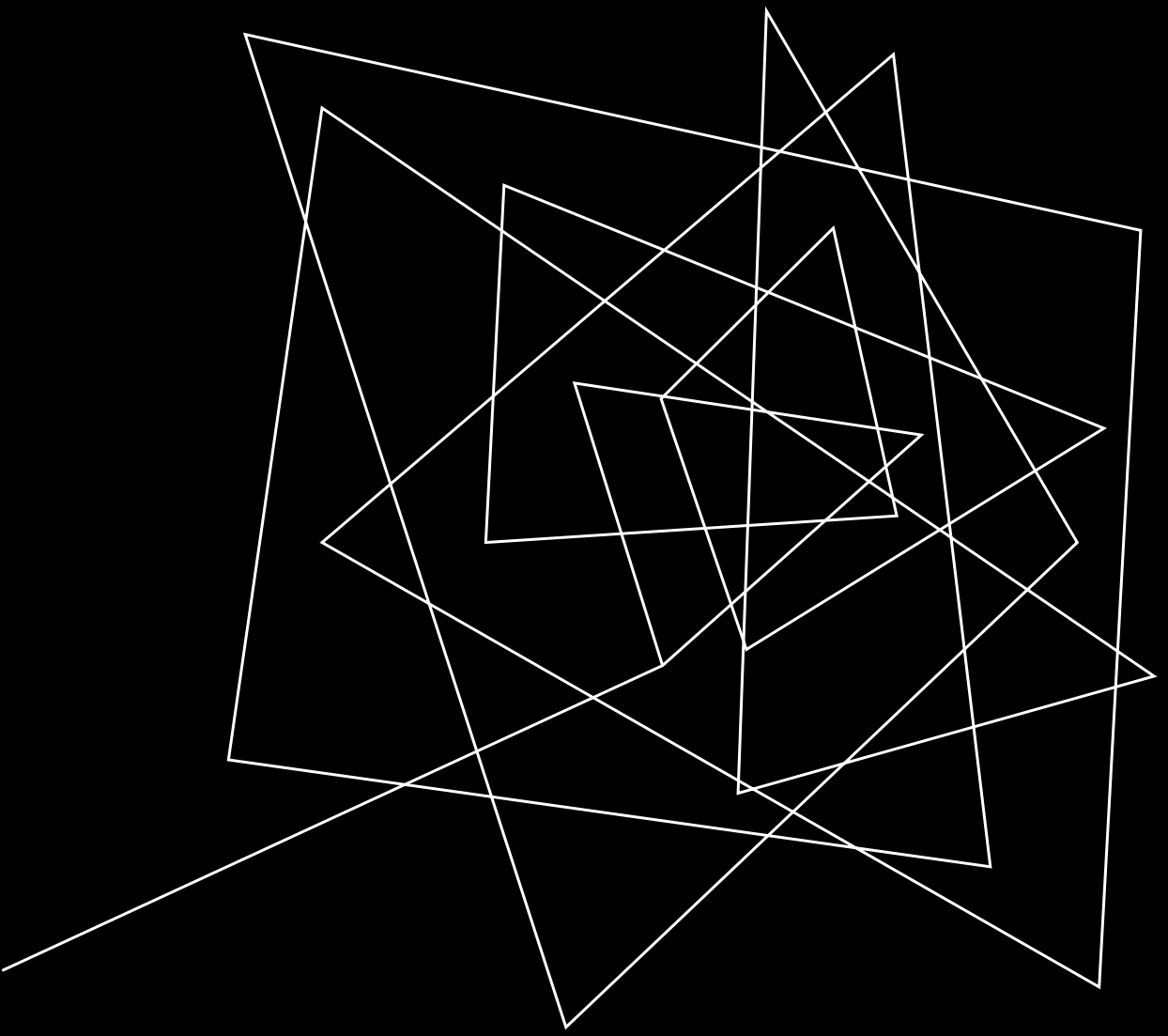
INTRODUCTION TO CHI SQUARED ANALYSIS

Chi squared Analysis or more precisely, Pearson's Chi-squared test, is a type of statistical analysis which is used to determine whether a significant correlation between two categorical variables exists or not.

It actually evaluates the likelihood of independence between the variables, which means we can evaluate the existence of relation but we can not explore further what kind of relation is that. Once we determine the relation, we can then apply other methods to recognize positive or negative relation.

For Chi squared test;

- Both the variable fields should belong to the same population.
- Variables should contain the categorical values, for example, YES/NO, Male/Female, True/False, etc.



PRIMARY GOAL

To determine the relationship between two categorical variables namely; Cylinders and Type in carsdatabase dataset.

ANALYSIS DESCRIPTION

Problem Statement: In the carsdatabase dataset, the Type column records the types of cars, whereas the Cylinders column records the number of cylinders along with the exception of an engine having rotary type. Categorical variable 'Type' has values; "Small", "Midsize", "Large", "Compact", "Sporty" and "Van". Again, in the variable 'Cylinders', allowed values are "3", "4", "5", "6", "8" and "rotary".

Test the hypothesis that Number of Cylinders in car engine is correlated with the Type of car sold.

Step 1: Stating Null and Alternate Hypothesis

Null Hypothesis H_o: Number of cylinders is independent of the car type

Alternate Hypothesis H_a: Number of cylinders depends on the car type

Step 2: Specifying the significance level

Significance level $\alpha = 0.05$

Step 3: Conducting Chi-squared test

```
        Pearson's Chi-squared test

data:  Tbl
X-squared = 78.935, df = 25, p-value = 1.674e-07

Warning message:
In chisq.test(Tbl) : Chi-squared approximation may be incorrect
```

Step 4: Here p- value of 1.674e-07 is infinitesimally smaller than the 0.05 confidence interval. That indicates the rejection of Null hypothesis.

INSIGHTS FROM THE ANALYSIS

Conclusion: From the above Chi-squared test, we have got the p-value of 1.674e-07 which is smaller than the significance level $\alpha = 0.05$. Therefore, we reject the Null hypothesis that the Number of cylinders is independent of car type.

Hence, **we accept the alternate hypothesis that Number of cylinders depends upon the car type**.

However, small cell values in the contingency table give the warning message stating that the 'chi-squared approximation may be incorrect' thereby indicating the need for the enhanced solution.

CONTINGENCY TABLE

	Compact	Large	Midsize	Small	Sporty	Van
3	0	0	0	3	0	0
4	15	0	7	18	8	1
5	0	0	1	0	0	1
6	1	7	12	0	4	7
8	0	4	2	0	1	0
rotary	0	0	0	0	1	0

ENHANCED SOLUTION

Enhanced solution can be obtained by combining 2 columns of the contingency table and applying chi squared test to the new table.

As we can see from the contingency table that several values are 0, hence, we are required to add 2 columns and check the chi- squared results as follows.

```
newT2 = cbind(Tbl[,"Compact"]+ Tbl[,"Small"], Tbl[,"Large"], Tbl[,"Midsize"], Tbl[,"Sporty"], Tbl[,"Van"])
print(newT2)
chisq.test(newT2)

newT3 = cbind(Tbl[,"Compact"] + Tbl[,"Large"], Tbl[,"Midsize"], Tbl[,"Small"], Tbl[,"Sporty"] , Tbl[,"Van"])
print(newT3)
chisq.test(newT3)

newT4 = cbind(Tbl[,"Compact"], Tbl[,"Large"]+Tbl[,"Midsize"], Tbl[,"Small"], Tbl[,"Sporty"] , Tbl[,"Van"])
print(newT4)
chisq.test(newT4)

newT5 = cbind(Tbl[,"Compact"], Tbl[,"Large"], Tbl[,"Midsize"]+ Tbl[,"Small"], Tbl[,"Sporty"] , Tbl[,"Van"])
print(newT5)
chisq.test(newT5)

newT6 = cbind(Tbl[,"Compact"], Tbl[,"Large"], Tbl[,"Midsize"], Tbl[,"Small"]+Tbl[,"Sporty"], Tbl[,"Van"])
print(newT6)
chisq.test(newT6)

newT7 = cbind(Tbl[,"Compact"] + Tbl[,"Van"], Tbl[,"Large"], Tbl[,"Midsize"], Tbl[,"Small"], Tbl[,"Sporty"])
print(newT7)
chisq.test(newT7)
```

EXPLANATION

However, all these combinations result in the same warning message.
Following images will show Chi squared test trial for 2 combinations;

```
> newT2 = cbind(Tbl[,"Compact"]+ Tbl[,"Small"], Tbl[,"Large"], Tbl[,"Midsize"], Tbl[,"Sporty"], Tbl[,"Van"])
> print(newT2)
       [,1] [,2] [,3] [,4] [,5]
3         3    0    0    0    0
4        33    0    7    8    1
5         0    0    1    0    1
6         1    7   12    4    7
8         0    4    2    1    0
rotary    0    0    0    1    0
> chisq.test(newT2)

        Pearson's Chi-squared test

data:  newT2
X-squared = 72.972, df = 20, p-value = 5.916e-08

Warning message:
In chisq.test(newT2) : Chi-squared approximation may be incorrect
```

```
> newT3 = cbind(Tbl[,"Compact"] + Tbl[,"Large"], Tbl[,"Midsize"], Tbl[,"Small"], Tbl[,"Sporty"] , Tbl[,"Van"])
> print(newT3)
       [,1] [,2] [,3] [,4] [,5]
3         0    0    3    0    0
4        15    7   18    8    1
5         0    1    0    0    1
6         8   12    0    4    7
8         4    2    0    1    0
rotary    0    0    0    1    0
> chisq.test(newT3)

        Pearson's Chi-squared test

data:  newT3
X-squared = 50.17, df = 20, p-value = 0.0002094

Warning message:
In chisq.test(newT3) : Chi-squared approximation may be incorrect
```

RECOMMENDATION OF OTHER TWO COMBINATIONS

In the contingency table, if we look at the row named ‘rotary’, most values associated are null, hence we can test by eliminating this row, which may be affecting the chi-squared test result.

Name the new dataset as carsdatabase_Copy.

```
> Tbl1 = table(carsdatabase_Copy$Cylinders, carsdatabase_Copy$Type)
> Tbl1

    Compact Large Midsize Small Sporty Van
  3       0     0       0     3      0   0
  4      15     0       7    18      8   1
  5       0     0       1     0      0   1
  6       1     7      12     0      4   7
  8       0     4       2     0      1   0
> # Chi-squared statistics
> chisq.test(Tbl1)

        Pearson's Chi-squared test

data:  Tbl1
X-squared = 72.508, df = 20, p-value = 7.057e-08

Warning message:
In chisq.test(Tbl1) : Chi-squared approximation may be incorrect
```

TWO COMBINATIONS AFTER ELIMINATING 'ROTARY' ROW

From the contingency table from carsdatabase_Copy, columns named 'Large' and 'Small' have 3 null values each and to ward off the warning message we try to combine these columns with any other column as follows and test the resultant table.

```
> newT_1 = cbind(Tbl1[,"Compact"], Tbl1[,"Large"]+Tbl1[,"Midsize"], Tbl1[,"Small"], Tbl1[,"Sporty"] , Tbl1[,"Van"])
> print(newT_1)
  [,1] [,2] [,3] [,4] [,5]
3    0    0    3    0    0
4   15    7   18    8    1
5    0    1    0    0    1
6    1   19    0    4    7
8    0    6    0    1    0
> chisq.test(newT_1)

        Pearson's Chi-squared test

data:  newT_1
X-squared = 63.068, df = 16, p-value = 1.579e-07

Warning message:
In chisq.test(newT_1) : Chi-squared approximation may be incorrect
> # Chi-squared test on new table
> newT_2 = cbind(Tbl1[,"Compact"], Tbl1[,"Large"],Tbl1[,"Midsize"], Tbl1[,"Small"]+Tbl1[,"Sporty"] , Tbl1[,"Van"])
> print(newT_2)
  [,1] [,2] [,3] [,4] [,5]
3    0    0    0    3    0
4   15    0    7   26    1
5    0    0    1    0    1
6    1    7   12    4    7
8    0    4    2    1    0
> chisq.test(newT_2)

        Pearson's Chi-squared test

data:  newT_2
X-squared = 63.721, df = 16, p-value = 1.221e-07

Warning message:
In chisq.test(newT_2) : Chi-squared approximation may be incorrect
```

However, in this case too, with these combinations, the warning message has not been eliminated.

Also, in all the combinations we get the p-value smaller than 0.05 significance level only.

If we try combining above two combinations and test, we again get the following output with warning message.

```
        Pearson's Chi-squared test

data:  newT_4
X-squared = 54.282, df = 12, p-value = 2.434e-07

Warning message:
In chisq.test(newT_4) : Chi-squared approximation may be incorrect
```

```
        Pearson's Chi-squared test

data:  newT_5
X-squared = 33.773, df = 12, p-value = 0.0007323

Warning message:
In chisq.test(newT_5) : Chi-squared approximation may be incorrect
```

R syntax

```
library(MASS)
View(carsdatabase_Copy$Cylinders)
View(carsdatabase_Copy$Type)

# Creating Table
Tbl1 = table(carsdatabase_Copy$Cylinders, carsdatabase_Copy$Type)
Tbl1
# Chi-squared statistics
chisq.test(Tbl1)


newT_1 = cbind(Tbl1[,"Compact"], Tbl1[,"Large"]+Tbl1[,"Midsize"], Tbl1[,"Small"], Tbl1[,"Sporty"] ,
Tbl1[,"Van"])
print(newT_1)
chisq.test(newT_1)

# Chi-squared test on new table
newT_2 = cbind(Tbl1[,"Compact"], Tbl1[,"Large"],Tbl1[,"Midsize"], Tbl1[,"Small"]+Tbl1[,"Sporty"] ,
Tbl1[,"Van"])
print(newT_2)
chisq.test(newT_2)


# Chi-squared test on new table
newT_3 = cbind(Tbl1[,"Compact"],Tbl1[,"Midsize"], Tbl1[,"Small"]+Tbl1[,"Sporty"] ,
Tbl1[,"Large"]+Tbl1[,"Van"])
print(newT_3)
chisq.test(newT_3)

# Chi-squared test on new table
newT_4 = cbind(Tbl1[,"Compact"], Tbl1[,"Large"]+Tbl1[,"Midsize"], Tbl1[,"Small"]+Tbl1[,"Sporty"] ,
Tbl1[,"Van"])
print(newT_4)
chisq.test(newT_4)

# Chi-squared test on new table
newT_5 = cbind(Tbl1[,"Compact"], Tbl1[,"Midsize"], Tbl1[,"Small"]+Tbl1[,"Sporty"]+Tbl1[,"Large"] ,
Tbl1[,"Van"])
print(newT_5)
chisq.test(newT_5)
```

CONCLUSION

Evidently, the likelihood of chi-squared approximation being incorrect implies that the test is inappropriate for the chosen dataset, which is carsdatabase. In this case, we can infer that alternate test of chi-squared test for the categorical variables should be considered.

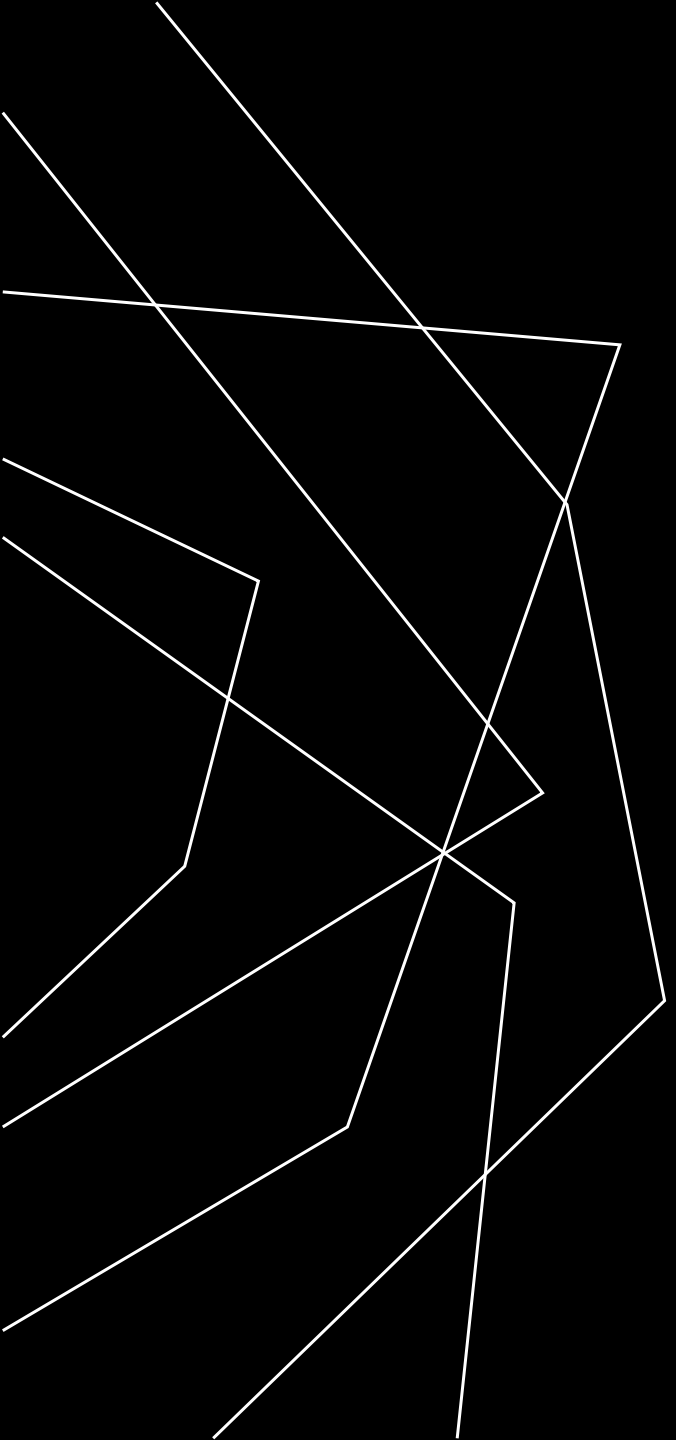
Fisher's Exact test can be used as an alternative to Chi-Squared test for our dataset.

R SYNTAX FOR THE CHI-SQUARED TEST

```
library(MASS)
View(carsdatabase$Cylinders)
View(carsdatabase$Type)

# Creating Table
Tbl = table(carsdatabase$Cylinders, carsdatabase$Type)
Tbl
# Chi-squared statistics
chisq.test(Tbl)

# Chi-squared test on new table
newT = cbind(Tbl[,"Compact"], Tbl[,"Large"], Tbl[,"Midsize"], Tbl[,"Small"],
Tbl[,"Sporty"] + Tbl[,"Van"])
print(newT)
chisq.test(newT)
```

R Syntax for different combinations

```
newT1 = cbind(Tbl[,"Compact"], Tbl[,"Midsize"], Tbl[,"Small"], Tbl[,"Sporty"]+Tbl[,"Large"],Tbl[,"Van"])
print(newT1)
chisq.test(newT1)

newT2 = cbind(Tbl[,"Compact"]+ Tbl[,"Small"], Tbl[,"Large"], Tbl[,"Midsize"], Tbl[,"Sporty"], Tbl[,"Van"])
print(newT2)
chisq.test(newT2)

newT3 = cbind(Tbl[,"Compact"] + Tbl[,"Large"], Tbl[,"Midsize"], Tbl[,"Small"], Tbl[,"Sporty"] , Tbl[,"Van"])
print(newT3)
chisq.test(newT3)

newT4 = cbind(Tbl[,"Compact"], Tbl[,"Large"]+Tbl[,"Midsize"], Tbl[,"Small"], Tbl[,"Sporty"] , Tbl[,"Van"])
print(newT4)
chisq.test(newT4)

newT5 = cbind(Tbl[,"Compact"], Tbl[,"Large"], Tbl[,"Midsize"]+ Tbl[,"Small"], Tbl[,"Sporty"] , Tbl[,"Van"])
print(newT5)
chisq.test(newT5)

newT6 = cbind(Tbl[,"Compact"], Tbl[,"Large"], Tbl[,"Midsize"], Tbl[,"Small"]+Tbl[,"Sporty"], Tbl[,"Van"])
print(newT6)
chisq.test(newT6)

newT7 = cbind(Tbl[,"Compact"] + Tbl[,"Van"], Tbl[,"Large"], Tbl[,"Midsize"], Tbl[,"Small"], Tbl[,"Sporty"])
print(newT7)
chisq.test(newT7)

newT8 = cbind(Tbl[,"Compact"], Tbl[,"Large"]+Tbl[,"Midsize"], Tbl[,"Small"]+ Tbl[,"Sporty"] + Tbl[,"Van"])
print(newT8)
chisq.test(newT8)
```



THANK YOU