Analytical Approach and Statistical Backing

25 July 2024 12:07

Scenario:

1. As of 2022, dragons are attacking the city causing havoc on the houses of city residents. In late 2022, city introduces two programs; Fire Control and Pest Prevention to reduce the dragon related risks on the houses. This is measured using DRS (Dragon Risk Score), the lower the safer.

2. The city has also introduced subsidy to support diverse population and households for the installation of dragon safety products.

a. Income<50k will receive complete (100%) subsidy
b. 50,000<= income <120k will receive 50% subsidy
c. Income>= will receive 25% subsidy

Assumption:

1. City is taking complete responsibility of waste management, hence the resident's share for the service installation expenses includes solely the cost of service and not waste management.

2. Other than the cost of each service, the residents weren't aware of the performance of each service including the ability of Pest Prevention and Fire Control to reduce the DRS score, the amount of waste generated, and the total time required to fully function the service.

Statistical Tests:

Why?

- To provide additional mathematical and statistical evidence on the data to support our decisions and explored insights.

What are we performing?

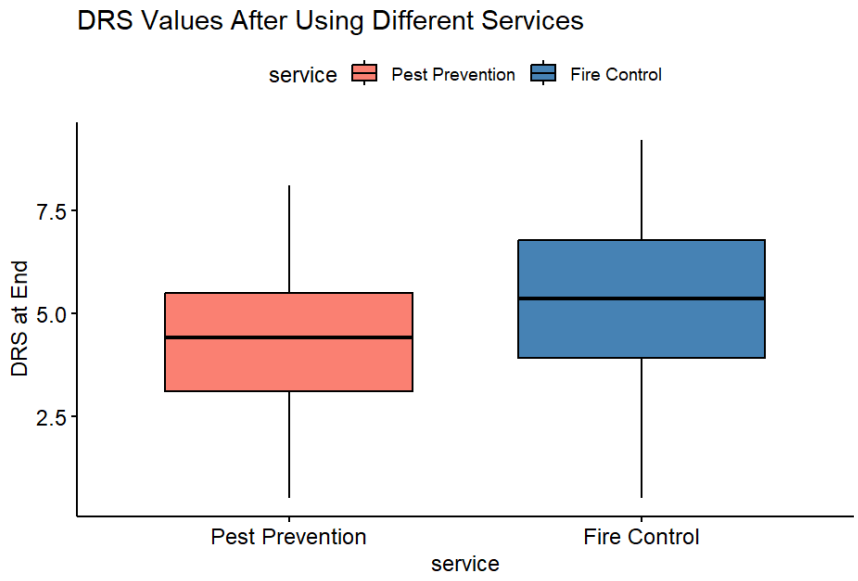
- Normality test: Checking if the distribution of numerical values are normally distributed. Performing Wilk-Shapiro test to analyze if the distribution is normal. If p-value>=0.05, the distribution is normal and we can proceed with parametric tests. If not, non-parametric tests.

- Parametric test: If the numerical data distribution is normal, we can find if there's significant difference within two numerical samples (granted both have passed Wilk-Shapiro test).

- Non-parametric test: If the numerical data is not normally distributed, i.e. p-value<0.05, we reject the null hypothesis and proceed to non-parametric test. This test does not assume and can handle numerical values/samples that aren't normally distributed.

Statistical Tests:

1. ***Checking if there's significant difference in the DRS at completion after using the two services.***



Summary of the distribution:

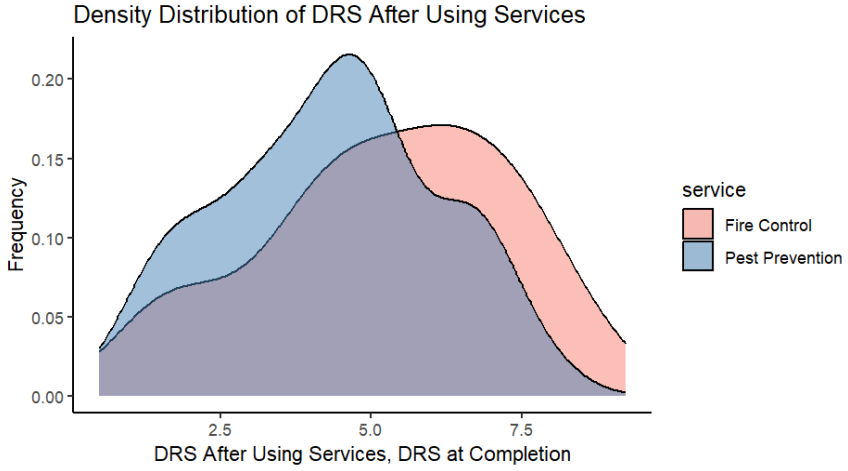
service	count	median	IQR
<chr>	<int>	<dbl>	<dbl>
1 Fire Control	99	5.36	2.87
2 Pest Prevention	119	4.4	2.4

a. The Median DRS at completion is higher for Fire Control than Pest Prevention service. This shows that that Pest Prevention is more effective in lowering the DRS than Fire Control.

Deduction:

There's a strong statistical evidence suggesting that the households using the Pest Prevention have lower median DRS values than Fire Control. The lower IQR value for Pest Prevention suggest that the middle 50% of the houses appear to have lower DRS score than the middle 50% of the houses using Fire Control.

Normality Test:
Checking the distribution of DRS after completion, if normal we can perform parametric test to analyze if there's significant difference between the DRS values of both services, else we will perform non-parametric test.



Wilk-Shapiro Test for normality:

1. p-value for pest prevention DRS at completion= 0.04345 < 0.05, we reject the null hypothesis and accept the alternative (the DRS at completion value for pest prevention is not normally distributed)

2. p-value for fire control DRS at completion= 0.0532 >0.05, we accept the null hypothesis that the DRS at completion value is normally distributed.

Since the p-values for the both services differ slightly from the significance value of 0.05, we could conduct parametric and non-parametric tests to assess the difference between DRS at completion values for both services.

Statistical Test:

1. Non Parametric: Wilcox Rank Sum/Mann-Whitney U test

   p-value= 0.0005153 < 0.05, we reject the null hypothesis (H0) and accept the alternative hypothesis (there's significant difference between the DRS distribution of both services.)

2. Parametric: Paired t-test

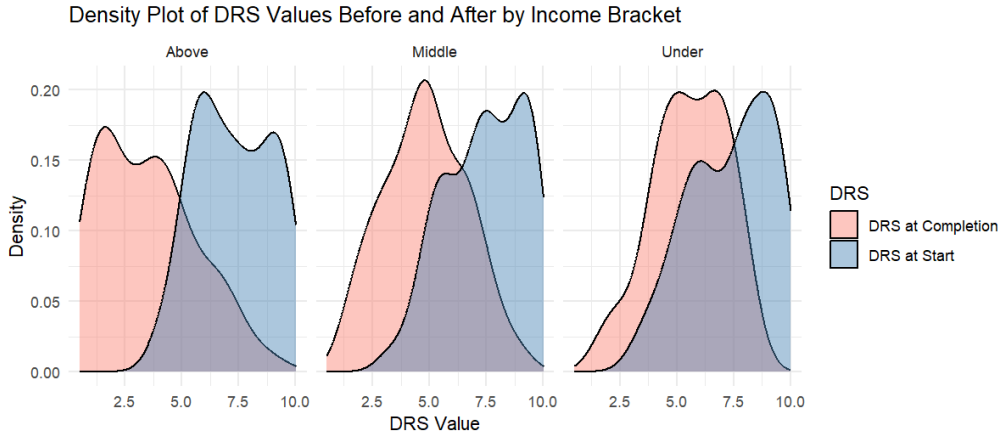
   p-value= 0.0006999 < 0.05, we reject the null hypothesis (H0) and accept the alternative hypothesis (there's significant difference between the mean DRS values of both services.)

Final Statement: For both the services, there's significant difference between DRS at completion values. We have strong statistical evidence suggesting differences in values which provide insight that pest prevention is effective than fire control.

---

2. *Analyzing if households by income brackets are safer than before*

Here we are looking at DRS scores before and after the completion of services by income brackets.



	income_bracket	variable	skewness	kurtosis
	<chr>	<chr>	<dbl>	<dbl>
1	Above	drs_at_completion	0.537	2.41
2	Above	drs_at_start	0.0179	1.80
3	Middle	drs_at_completion	0.0749	2.37
4	Middle	drs_at_start	-0.348	2.13
5	Under	drs_at_completion	-0.366	2.49
6	Under	drs_at_start	-0.381	2.05

The density plot shows the DRS distribution before and after installation of dragon services for 3 income brackets. The income brackets are divided into three categories based on the conditions provided by for subsidy plans.

Skewness is the measure of distortion or shift of density plot from the standard normal distribution. If the density plot has peak(s) tending towards 0 or origin, it's called positive skewness and the opposite is called negative skewness.

For the above category as a collective, has seen s strong distortion of skewness when the DRS scores for before and after are compared. On the x-axis, the DRS scores are presented, if the peak for DRS at completion is towards the origin or 0, this means that a significant amount of households with an income bracket tend to have smaller DRS values,

meaning they are more safer.

A significant shift has been noticed in the following order:
Above (0.5191) > Middle (0.4229) >Under (0.015)

Normality Test:
To check if the DRS value distribution for all the households.

	income_bracket	DRS_before_p.value	DRS_after_p.value
	<chr>	<dbl>	<dbl>
1	Above	0.0323	0.0204
2	Middle	0.000649	0.354
3	Under	0.00931	0.234

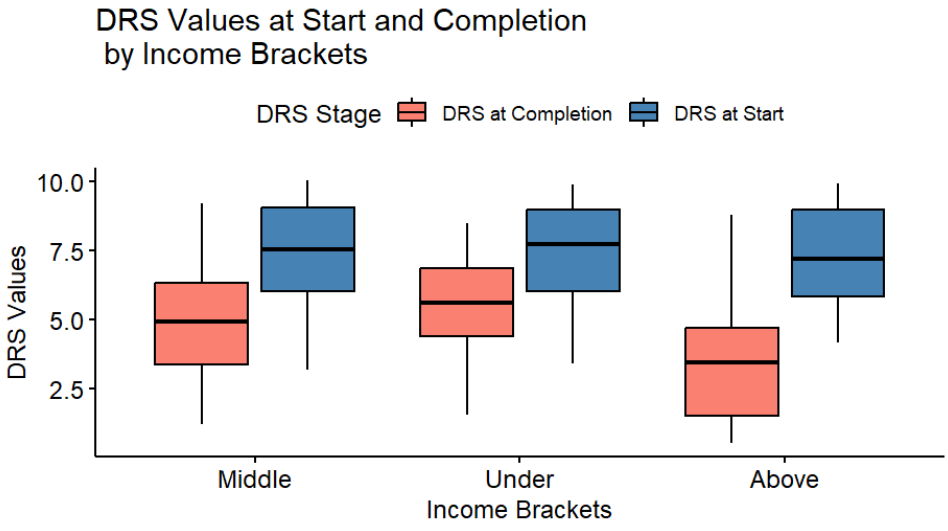
By performing Wilk-Shapiro test and p-values, it's evident that the DRS values for Middle and Under follow a normal distribution while the Above bracket doesn't.

Non-Parametric Test:
The p values generated for all the income categories were < 0.05 significance level, indicating significant differences noticed in the DRS values before and after the installation of the dragon services.

P-value (above):
```
cannot compute exact p-value with ties
```

P-value (middle): p<2.2e-16

P-value (under): p<2.45e^-11

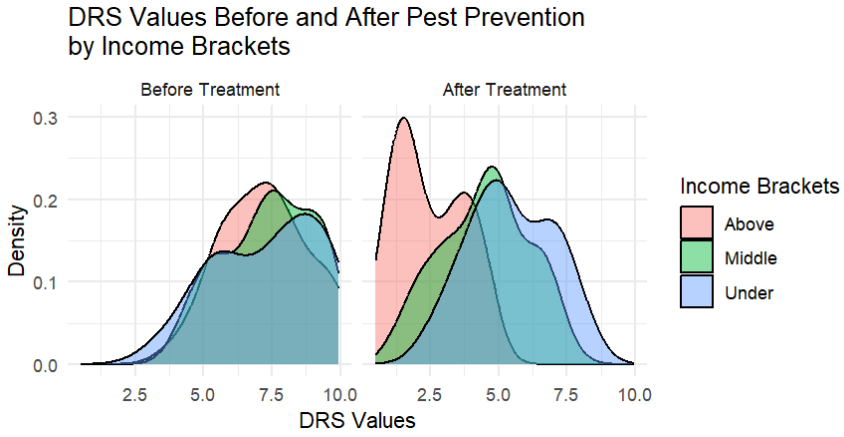


	income_bracket	count	median_DRS_start	median_DRS_completion	IQR_DRS_start	IQR_DRS_completion
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
1	Above	46	7.2	3.44	3.12	3.18
2	Middle	113	7.55	4.9	3.06	2.94
3	Under	59	7.71	5.6	2.94	2.48

Through the box plot, it's evident that the DRS values are decreasing after the installation and the median values are dropping for all brackets.
The DRS values are sparse for above category, meanwhile the middle 50% of the DRS values for middle and under are concentrated near their respective median values.

Final Statement: All the brackets show decrease in the median DRS values. From the above-shown diagrams, statistical tests and their respective p-values show validate to reject the null hypothesis and accept that significant differences in the DRS values for all the brackets were observed.

3. *Statistical Visual Analysis of DRS change by Income Bracket for Services*

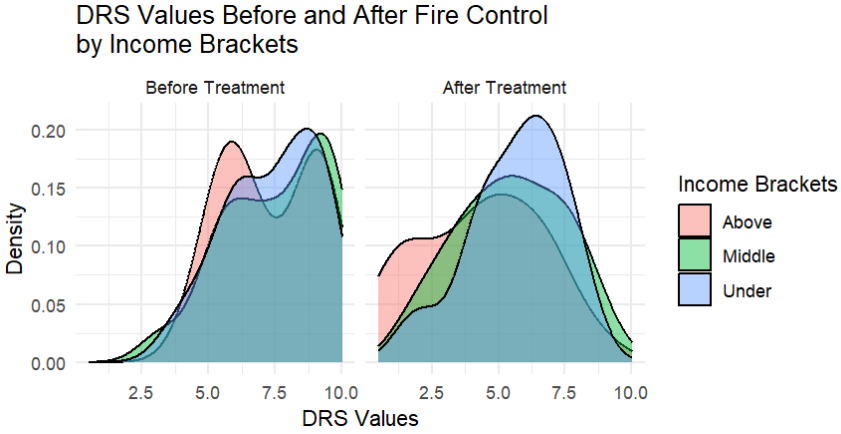


income_bracket	variable	skewness
<chr>	<fct>	<dbl>
1 Above	drs_at_start	0.0504
2 Above	drs_at_completion	0.248
3 Middle	drs_at_start	-0.254
4 Middle	drs_at_completion	-0.126
5 Under	drs_at_start	-0.338
6 Under	drs_at_completion	-0.0102

Similar to the previous versions of statistical test, this test shows us evidence of improved skewness toward positive side, indicating improvement in the DRS values of households with Pest Prevention service.

The improvement in witnessed though the difference of skewness values. The order can be determined by the values: Under (0.3278) > Above (0.1976) > Middle (0.128)

Under income brackets are benefited the most from using the Pest Prevention method/service.



income_bracket	variable	skewness
<chr>	<fct>	<dbl>
1 Above	drs_at_start	-0.00865
2 Above	drs_at_completion	-0.0380
3 Middle	drs_at_start	-0.492
4 Middle	drs_at_completion	-0.122
5 Under	drs_at_start	-0.411
6 Under	drs_at_completion	-0.595

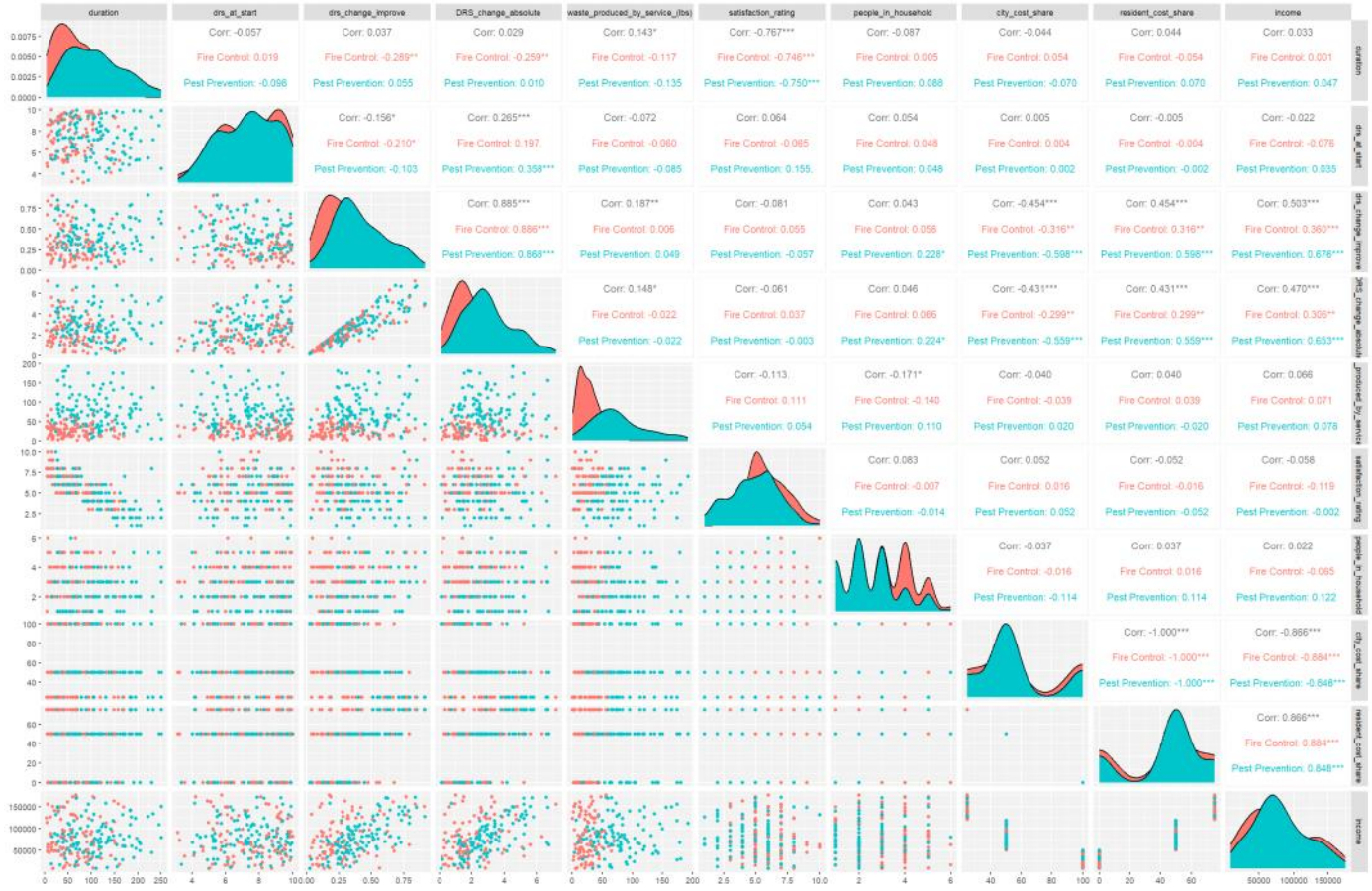
The skewness change is not significant as in the Pest Prevention method. The improvement in the DRS values are improving however, not as significantly as Pest Prevention.

The order can be determined by the values: Middle (0.37) > Above (0.0293) > Under (-0.184)

The skewness difference has shown improvement in the middle bracket, however, the above bracket has benefitted slightly and the DRS values have been moderately less. Meanwhile, the under bracket is not benefitting from this service, a negative skew difference, meaning more negative skewness tends to make the density left skewed, away from the ideal or the necessary condition/scenario.

---

4. Assessing correlations amongst numerical and categorical variables.



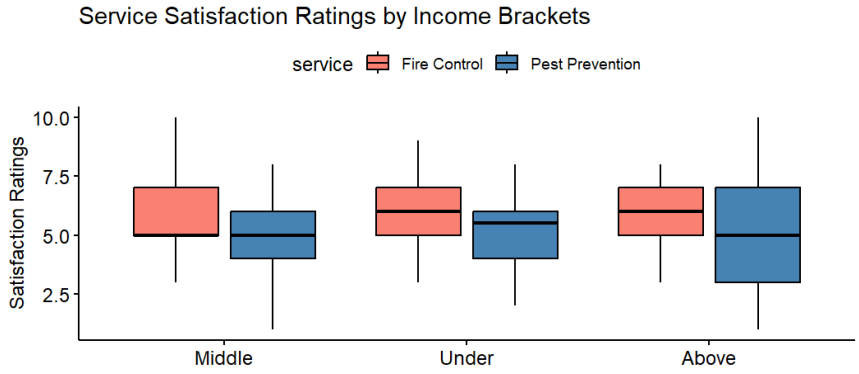
The "Corr" scores/coefficients in the correlation matrix provide us strength and signs provide us direction between numerical variables.
The coefficient ranges from -1 to 1. The signs positive (+) and negative (-) depicts directly and inversely proportional relationship between variables respectively.

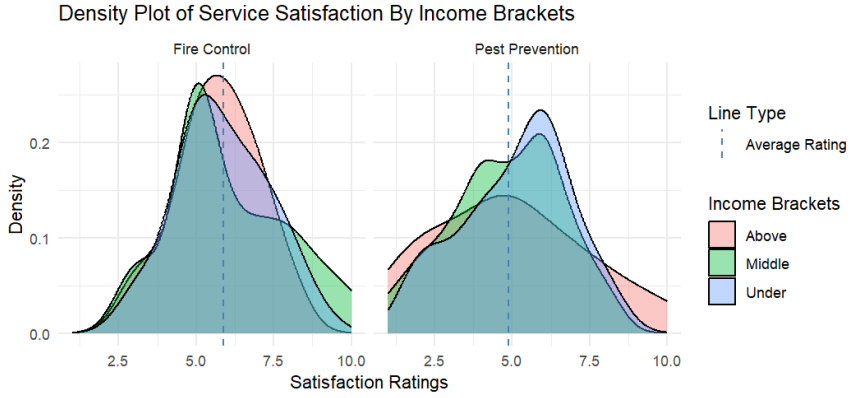
A positive coefficient means the variables are positively correlated to each other, increase in one variable is proportional to the increase in the corresponding correlated variable.
A negative coefficient means the variables are negatively correlate to each other, an increase in one variable is proportional to the decrease in another variable.

The asterisk (*) symbols represent the statistical significance of the correlational relationship. This suggests that the relationship is unlikely to have caused by random chance.

The higher number of asterisk suggests higher significance.

Additional visualizations to support analysis.





service	income_bracket	skewness	kurtosis
<chr>	<chr>	<dbl>	<dbl>
1 Fire Control	Above	-0.235	2.57
2 Fire Control	Middle	0.392	2.41
3 Fire Control	Under	0.0882	2.45
4 Pest Prevention	Above	0.316	2.32
5 Pest Prevention	Middle	-0.240	2.35
6 Pest Prevention	Under	-0.316	2.26

service	skewness	kurtosis
<chr>	<dbl>	<dbl>
1 Fire Control	0.312	2.76
2 Pest Prevention	-0.0585	2.51