

HOUSE PRICE

Introduction

In this data set we will be investigating empirically the variety of possible factors that influenced the price of housing using the linear regression approach. There are numerous factors in this data like:

Grade: which is the possible remark or review from individuals

Sqft_lot: which is the size of the room.

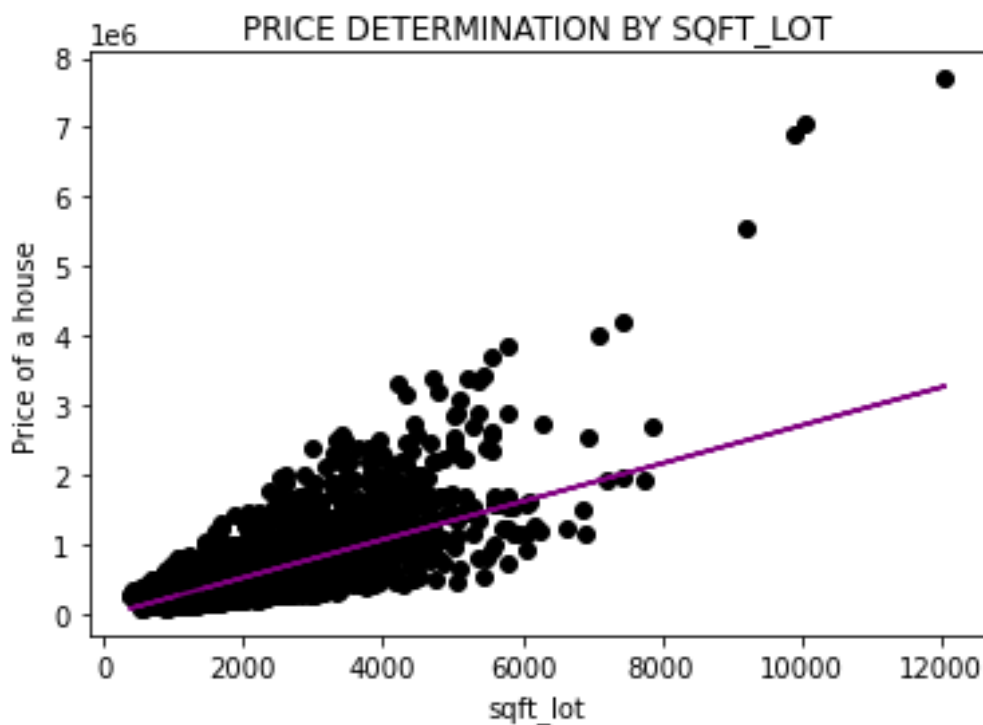
Zip code: this determines the location of the apartment. just to mention but few, for the purpose of this research we will consider variable(s) with highest possible correlation / coefficient.

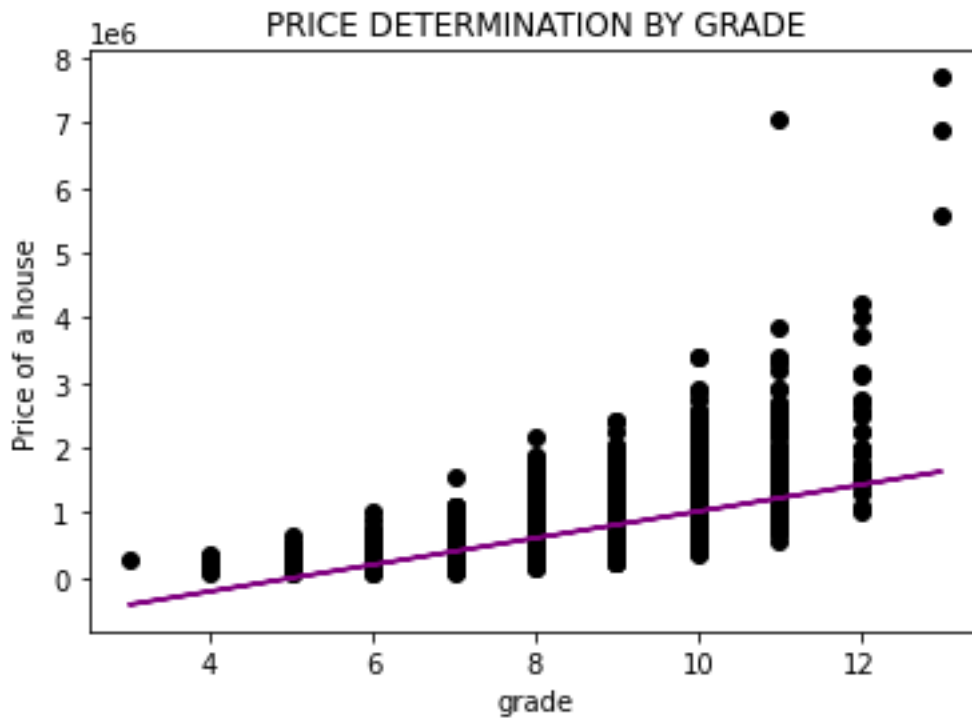
We started by importing all the required library command for analysing our regression model, then we proceeded to importing our house price data set into python, we proceeded to confirming our data by printing the head, then proceeded to viewing the general properties of our data using descriptive statistics, we then test for correlation to chose the best independent variables that best impact price determination.

Visualization With Single Variable (SQFT_LOT & GRADE)

Where x = independent variable

Y = target variable

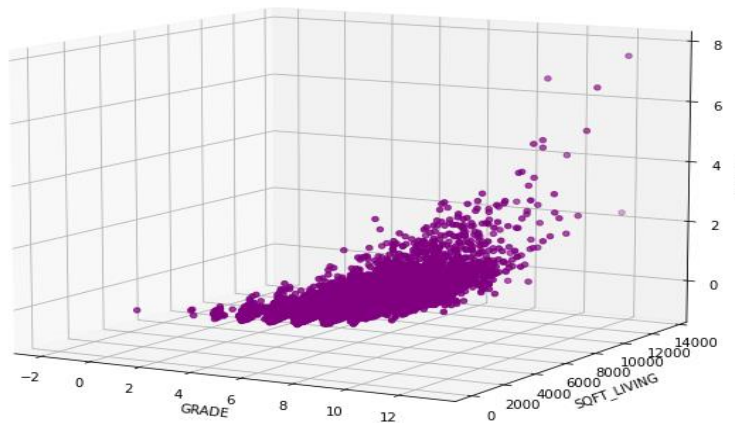




The scatter plot visualization illustrates a positive relation using sqft_lot and grade independently as one of the factors using the linear regression model, the mean squared error between the actual and predicted value appears a bit with the co-efficient of determination at equilibrium, it can be significantly justified that the condition and size of a house plays a huge role in price determination, larger apartments with wide rooms are generally more expensive and more valuable with less repair, this metric measures the cost of property in relation to its size, these factor attracts prospective investors to compare properties with ranges of sizes and features to determine the relative value of each.

Multiple Variables: Grade and Sqft-lot

Using a 3D plot with 3 variables, x and y are independent variables while z is the target variable



These investigations further explore the positive correlation between two variables (Grade and Sqft_lot) in determining the price of a house. Grade typically explains the condition of the property while sqft_lot generally refers to the overall size of the property which determines the number of rooms (kitchen, toilet, pantry, living room, study room, bedrooms) the size of house can accommodate. The data set shows a moderate positive influence on price determination and will lower our regression model's error component.

We then proceed to training and splitting our data & model, our analysis reveals that the test performs somewhat better than the train.

The Results are given below

Coefficient: [98680.60889071 177.90721499]

Intercept: -584908.8045087642

Mean squared error: -8.149072527885437e-10

Co-efficient of determination: 0.54

According to our regression co-efficient the price of a house is predicted to increase by 986.68 when grade increases by 1, increases by 177 when sqft_lot increases by 1. The intercept is the predicted value of y when x is 0, i.e. the point where the function crosses y axis.

Given that the mean squared error is a bit high, this model will almost certainly have low prediction accuracy. Our coefficient of determination is 0.54 which means that the relationship between price, grade, and sqft_lot is quite weak, because the higher R² the better regression. The more features the better our model performs as evidenced below:

Coefficients: [1.02569413e+05 1.83474227e+02 -2.67439032e+04 -1.51028344e+04
-2.62159983e-01 -8.17981626e+03 6.13311569e+05 6.31587507e+04
5.77491631e+04]

Intercept: -704716.8237396934

Mean squared error: 58428222317.14476013

Coefficient of determination: 0.60

Improvement

The model can yield better results:

By rescaling the data set to normalise given features

By using correlative word to describe how more independent variables affect output variable

By invoking polynomial terms express the non-linear relation.

In summary, Properties with elevated grade constantly grip high price, symbolizing that consumers are willing to offer premium for well rated and conserved house in relation with the sqft_lot.

Critics: The data could have been less flawed if the factors below were considered while collating the data as they influenced the value of property:

Economic Condition: the situation of the economy can also affect the price of a house, when the economy is stable and people possess more disposable income, they become more willing to spend on housing, which can uplift the demand and prices simultaneously,

Home Maintenance: The condition of a house can also impact its price, house with modern facilities, and desirable furniture and fittings tend to be more lucrative than houses with outdated fittings.

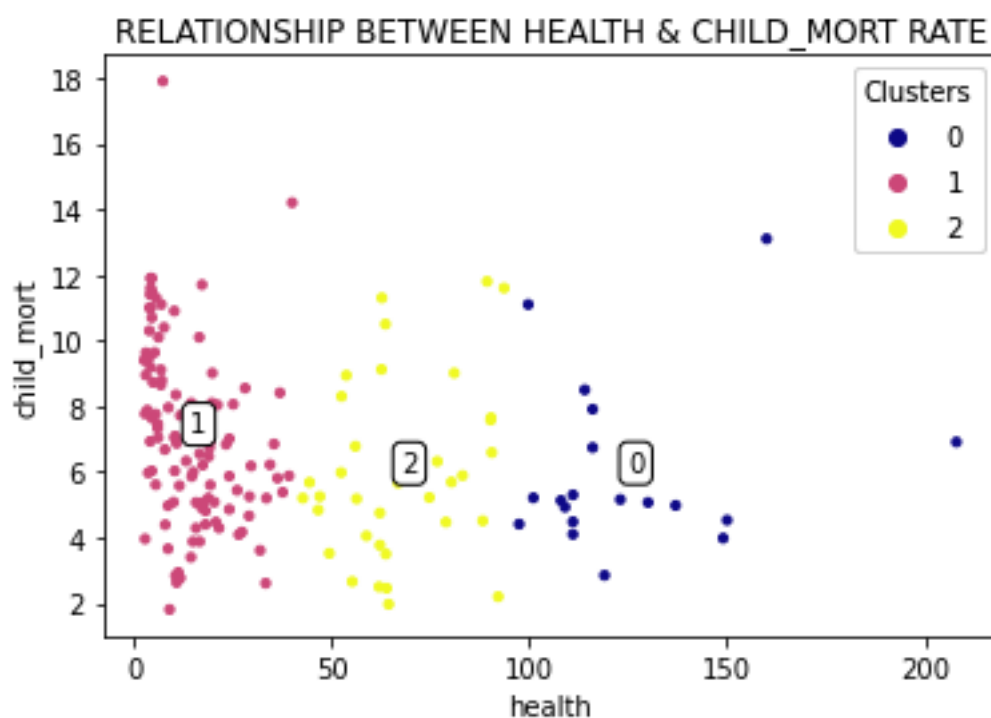
In summary the above factors, location, low crime rate, air and noise pollution, property maintenance, and access to basic amenities (shopping mall, public transport, school) determines the pricing of a house. In addition, while the size of a house plays a significant role in price determination the grade value and other factors also have an impact in price determination and purchase attraction in considering the value of a property.

COUNTRY DATA

CLUSTERING

In this dataset we will be communicating some specific findings using country data with specific information on child mortality, Exports, Health, Imports, Income, Inflation, Life Expectancy, Total_fert, Gdpp using KMeans and Mean shift Algorithm and using cluster in grouping similar datasets together.

1 KMeans

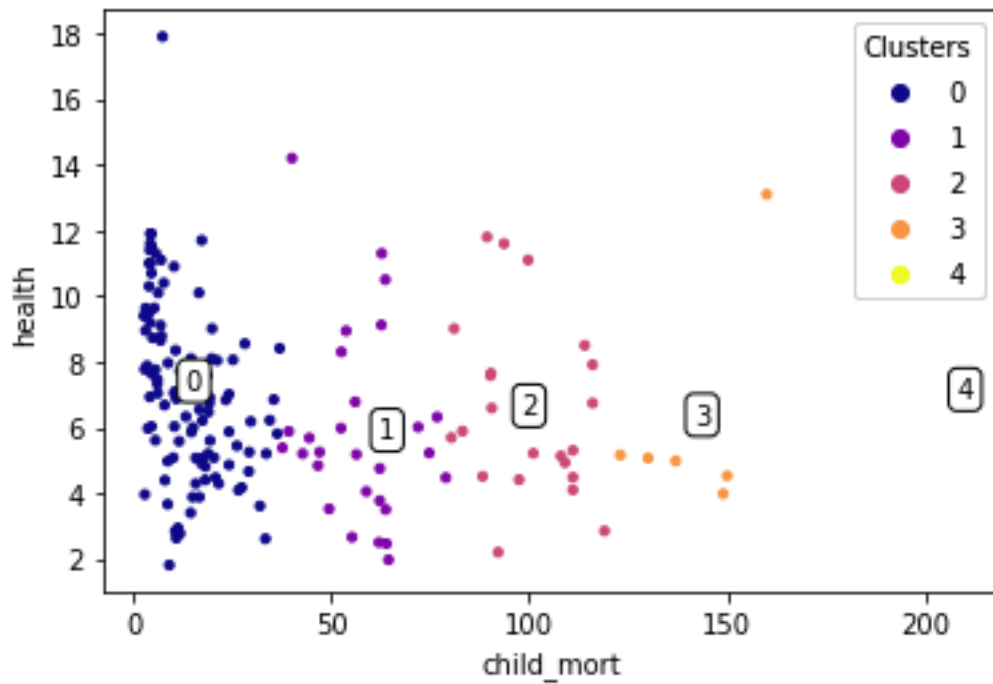


Centroids: $\begin{bmatrix} 124.74210526 & 6.02052632 \end{bmatrix}$

$\begin{bmatrix} 14.20625 & 7.20875 \end{bmatrix}$

$\begin{bmatrix} 67.49722222 & 6.0125 \end{bmatrix}$

2 Mean Shift



Centroids: [[13.09439252 7.17476636]

[61.92068966 5.67310345]

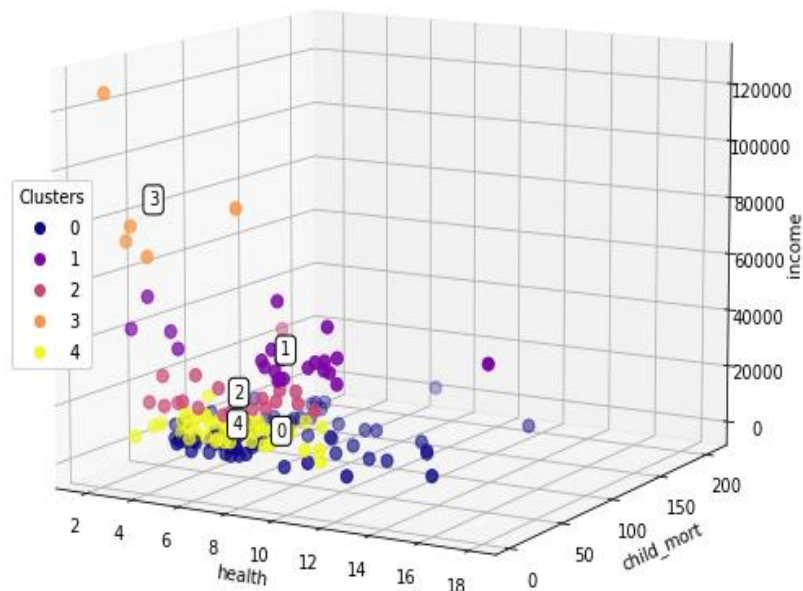
[97.80416667 6.41541667]

[141.5 6.13666667]

[208. 6.91]]

For the research, we will concentrate on the inverse relationship between health and child mortality then we take it further by considering additional variable which is the Income level.

KMeans 3 Variables



Centroids: $[[6.20625000e+00 \ 6.92041667e+01 \ 3.46990278e+03]$

$[9.49681818e+00 \ 5.50000000e+00 \ 4.38727273e+04]$

$[7.18391304e+00 \ 1.19043478e+01 \ 2.62608696e+04]$

$[3.80200000e+00 \ 7.18000000e+00 \ 8.89200000e+04]$

$[6.62666667e+00 \ 2.17266667e+01 \ 1.33228889e+04]]$

When additional features are added, the centroids are closer to one another and the clusters get extremely compact and crowded.

RESEARCH: Child mortality and health share a profound and intricate relation, establishing a focal aspect of world health discourse. Health is a cornerstone of human development and well-being, and its influence is intensely pronounced when considering child mortality. The correlation between health and child mortality is multifaceted and composite, as the physical, social and economic dimensions of health play crucial roles in determining the survival and thriving of young lives.

NBA_ROOKIE DATA

In this dataset I will apply a number of Machine learning models to classify the players accordingly :these algorithm are as follow

1 logistic regression

2 Gaussian Naïve-Bayes

3 Artificial Neural Networks

I will illustrate how each model perform

NEURAL NETWORKS

Some specific features with strong correlation were selected to test for accuracy and the number of mislabelled points with neurons and hidden layers , interestingly as the variable with the highest correlation and accuracy with the targets is Games Played, and the model is clearly a Binary Classification.

Features	Accuracy	Mislabelled points
1	0.68	168
2	0.62	150
3	0.66	150
4	0.65	155
10	0.67	148

This model was tested with numerous features(Games Played :Minutes Played :Points Per Game :Field Goals Made :Free Throw Made)

```
37
38
39 scaler = StandardScaler()
40 X_train = scaler.fit_transform(X_train)
41 X_test = scaler.transform(X_test)
42
43
44 # setup the neural network architecture
45 mlp = MLPClassifier(hidden_layer_sizes=(10,50,20),
46 activation="logistic", random_state=0, max_iter=2000)
47
48 mlp.fit(X_train, y_train)
49
50 print('value prediction %d:' %mlp.predict([[15,5,3,8,2]]))
51
52 # performance metrics
53 print('Our Accuracy is %.2f' % mlp.score(X_test, y_test))
54
```

	Name	Games Played	Minutes Played	...	Blocks	Turnovers	TARGET_5Yrs
0	128	36	27.4	...	0.4	1.3	0
1	45	35	26.9	...	0.5	1.6	0
2	521	74	15.3	...	0.3	1.0	0
3	820	58	11.6	...	0.1	1.0	1
4	871	48	11.5	...	0.4	0.8	1

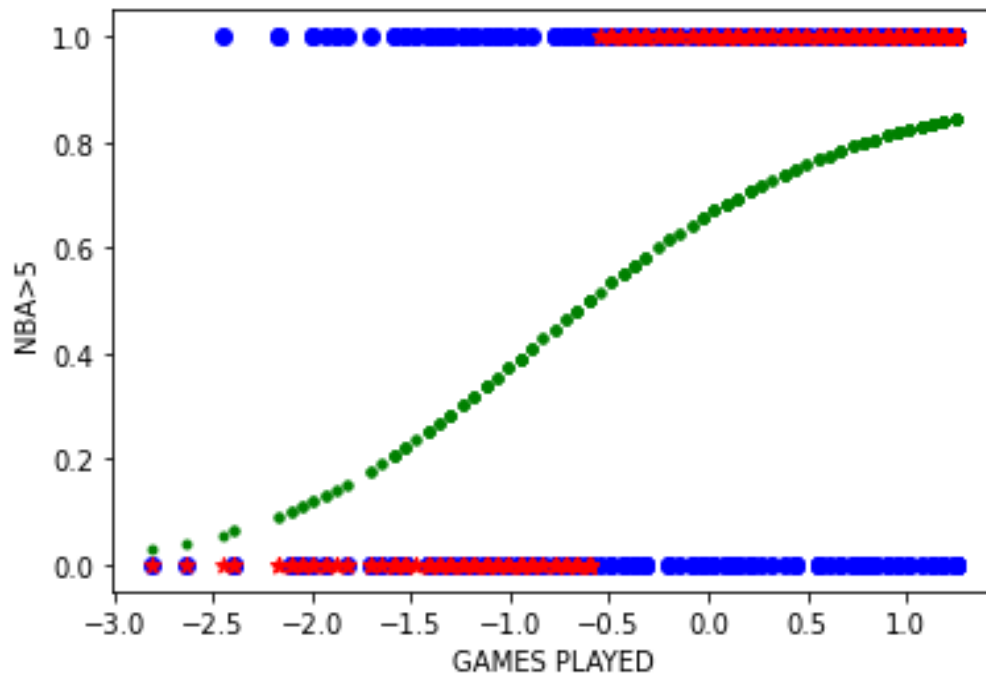
[5 rows x 21 columns]

Predict a value: [1]
value prediction 1:
Our Accuracy is 0.68
Number of mislabeled points out of a total 447 points : 144
gnb Predict value: [0]
Number of mislabeled points out of a total of 447 points: 144
Our accuracy is 0.68:

Python Console History

The above model using five features predicted with an accuracy of 0.68%, meaning the players with the same data as in Neural Networks would last <5 years in NBA.

GAUSSIAN NAÏVE BAYES

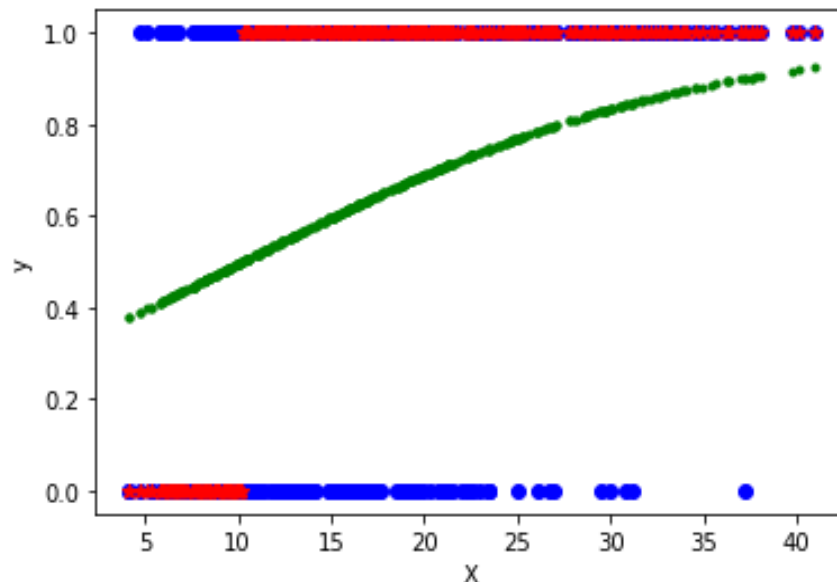


Features	Accuracy	Mislabelled points
1	0.68	168
2	0.63	167
3	0.62	168
4	0.63	166
10	0.62	168

The GNB predicted 68% accuracy, that the players with the same data as NN would last up to 5 years in the NBA

LOGISTIC REGRESSION

ACCURACY is 67%, With an accuracy of 67%, players with the same dataset as neural networks could not compete in the NBA for more than five years.



REPORT:

All our models as been tried on all our randomly generated model and they have appeared similar with constant parameters as there was no significant change in prediction, accuracy and mislabelled points, with their accuracy ranging between 0.62 – 0.68%. the model can be better improved if more features can be generated from existing features to aid prediction ability as each of them are unique in their own way , the GNB and Neural Network must have considered non-linear features which made them perform better and stable, compare to logistic regression that seems more vigorous and might have failed to converge. I would consider options like Mutual information for in-depth as an attempt to correlation algorithm didn't do much justice, tuning model hyperparameters, class balance optimization and eradication of redundant features could have improved the model.

ETHICS OF AI

The ethics of AI is an intricate and complicated topic that has been the central of attention in debate in recent time. The Trolley Problem is a test practice to consider if an action with less effect is worth implementing or an inaction with severe impact. Considering the context of autonomous vehicles, the product of technology should be programmed to make rational decision in a situation where destruction is inevitable. In the case of autonomous vehicle here the car makes decision without human intervention which can be harmful to many, based on automatic default. In a 1967 philosophy paper, Philippa Foot posed the issue of developing a general principle that may explain the divergent opinions that emerged in many versions of the narrative. Judith Jarvis Thomson coined the term "THE TROLLEY PROBLEM" in a 1976 essay that sparked a significant body of work.

According to Foot she pointed out, the doctrine of double effect is susceptible to counterexample if it is overly defined as the rule that morally acceptable actions with predictable negative effects are acceptable if those effects are not direct results i.e., if they are intended only indirectly. I will sight a lay man understanding to aid further understanding, sellers of harmful substance that they know will kill a large number of innocent people as fresh juice are not excluded from accountability because their primary goal is to maximise profit, not to intentionally cause death of their consumers. Even though the death of their customers are only an incidental consequence of their primary goal. However, the distinction between directly and indirectly intended consequences should be taken seriously because it helps to distinguish between situations where it would be morally tolerated if not required to carry out an action that one knows will result in the death of an innocent person and instances where it would be manifestly immoral to do so.

My understanding of the trolley problems of the context of autonomous vehicles causes a nuanced viewpoint. AI Algorithms, unlike human drivers must be pre-programmed with swift decision-making ethics that instructs how the vehicle should respond in emergency situations. This raises concerns about Programmers accountability, the moral values embedded in algorithms and the implications of these decisions on human lives. As artificial intelligence continues to address the challenges posed by the trolley problem in ensuring the future of vehicles, addressing the ethical challenges is crucial to ensuring the accountable deployment of autonomous vehicles in a globalized and diverse world.

REFERENCES

Philippa Foot, "The Problem of Abortion and The Doctrine of the Double Effect " in virtues and vices (Oxford Review, Number 5, 1967.)

Judith Jarvis Thomson, Killing , Letting Die , and The Trolley Problem, 59 The Monist 204-17 (1976)

Berkey, B 2019. Autonomous Vehicles, Business Ethics, and Risk Distribution in Hybrid Traffic: Autonomous Vehicle Ethics: Beyond The Trolley Problem(OUP)