International Journal of Humanities Social Science and Management (IJHSSM)
Volume 4, Issue 2, Mar.-Apr., 2024, pp: 1115-1122 www.ijhssm.org

Statistical Analysis of Used Cars

Lekh Raheja, *NMIMS University, Mumbai, Maharashtra.*
Mahi Nair, *NMIMS University, Mumbai, Maharashtra.*
Manav Rohra, *NMIMS University, Mumbai, Maharashtra.*
Mimansa Agarwal, *NMIMS University, Mumbai, Maharashtra.*
Ms. Tejaswini Angre*: Professor at NMIMS University, Mumbai, Maharashtra*
Corresponding Author: Maria Soni, NMIMS University, Mumbai, Maharashtra.

Date of Submission: 02-04-2024 Date of Acceptance: 12-04-2024

ABSTRACT
The used car market plays a significant role in the automotive industry, offering consumers an affordable alternative to new vehicles. This study employs statistical techniques to analyze secondary data on used cars, aiming to gain insights into usability, performance, and affordability factors that influence consumer decisions. The analysis utilizes time series methods, including moving averages and the least squares method, to identify trends and construct trendlines for variables such as mileage and engine displacement. Additionally, chain-based index numbers are computed to track relative changes in these variables over time, facilitating comparisons across periods. The findings reveal fluctuating patterns in mileage and engine displacement, potentially influenced by economic conditions, consumer preferences, and technological advancements. The trendline analysis suggests a decreasing trend in mileage, which may indicate evolving consumer demand for used cars. Furthermore, the index numbers highlight notable spikes and dips in mileage and engine displacement, offering valuable insights into market dynamics. This study contributes to a better understanding of the used car market, providing a quantitative basis for stakeholders to make informed decisions regarding pricing, inventory management, and consumer targeting strategies within the automotive industry.

KEYWORDS
Statistics, Descriptive Statistics, Times Series, Moving Average, Regression, Index Numbers.

I. INTRODUCTION

When we talk about purchasing a car, we usually think of buying a car first hand from a showroom i.e., completely new. The idea of even purchasing a used or used car is usually looked down upon irrespective of the car's performance statistics. Owning a car whether new or used can be of utmost benefit to anyone as it reduces the burden of day-to-day commuting, it also helps in increasing Mobility on a larger scale. However, it does have some of its downsides as well, such as High maintenance costs, rising fuel prices, parking issues etc. Our group has put forth secondary data of used cars, to analyse its usability, performance, and affordability which in turn would help determine whether used cars are worth purchasing or not. In our data, the independent variable is the year in which the car was manufactured, and our other dependant variables include the following: Engine displacement, Mileage and Price of the car. These 3 variables help a customer in determining, whether a car is worth purchasing or not. To build a further understanding of the data, we analysed it by the means of Time series trends using the moving average and least square method through which we obtained a trend line. We also used chain-based Index numbers to further analyse the predicted engine displacement and mileage of the cars and compare it to the past. This helps in drawing various conclusions regarding the overall and expected performance of the cars. Our analysis was conducted by the means of using an excel sheet to conduct the above-mentioned experiments.

II. DATASET

Since out dataset consists of more than 500 rows, we have attached a hyperlink that will direct the reader to our dataset.

QT project 26-30 BBA-I link.xlsx

The columns that we have selected for our analysis are: **Manufacturing year** and **Mileage** and **Engine Displacement.**

International Journal of Humanities Social Science and Management (IJHSSM)
Volume 4, Issue 2, Mar.-Apr., 2024, pp: 1115-1122 www.ijhssm.org

III. OBJECTIVE

This group project enables us to analyse and interpret the data through various statistical tools such as descriptive statistics, moving averages, and index numbers, and formulating trend lines by applying the least square method.

This study provides insight into the world of used cars and helps us analyse the factors which people consider while purchasing used cars. The practical applications of statistical tools like moving averages and index numbers help us to estimate future data and obtain the trend line for preference for used cars among the masses.

The project enables students to understand the application of the various statistical techniques using MS Excel which is widely used in many different industries and hands-on experience with Data Analysis tools in MS Excel.

IV. DESCRIPTIVE STATISTICS

Descriptive statistics involve methods and techniques used to summarize and describe the key features of a dataset. This includes measures such as mean, median, mode, range, variance, and standard deviation, providing insights into central tendency, variability, and distribution of the data without making inferences or conclusions beyond the dataset itself.

For our project and research purpose, descriptive statistics was of importance as it helped us find the Coefficient of Variation using the Mean and the Standard Deviation from the given table.

Mileage	
Mean	133812.553
Standard Error	2354.147099
Median	135776.5
Mode	0
Standard Deviation	75111.68162
Sample Variance	5641764716
Kurtosis	-0.363856414
Skewness	0.116925484
Range	398000
Minimum	0
Maximum	398000
Sum	136221179
Count	1018

The table given shows the Descriptive Statistics of the data selected by us. Now using the values of Mean and Standard Deviation, we have found the Coefficient of Variation.

$$\textit{Coefficient of variation} = \frac{\textit{Standard Deviation}}{\textit{mean}} \times 100$$

$$\frac{75111.68162}{133812.513} \times 100 = 56.14$$

$$= 56.14\%$$

The coefficient of variation (CV) is a statistical measure used to assess the relative variability of a dataset. The result is expressed as a percentage, in this case 56.14%.

From the Coefficient of Variation, we can conclude that the data is inconsistent as it is above 50%. This is because mileage, the dependant variable fluctuates with time and therefore shows big deviations in values, making the data inconsistent. From this we can conclude that the mileages exhibit moderate to high relative variability compared to their mean. However, it is important to note that 56% is close to the 50% cut off for deciding consistency or inconsistency therefore, while the data is inconsistent it may not be to a high degree.

V. MOVING AVERAGES

Moving average is a statistical method that smooths out data fluctuations by averaging consecutive data points within a specified window. It is commonly used in time series analysis to identify trends or patterns. The average is recalculated as new data becomes available by shifting the window forward.

Diverse types of moving averages are as follows:

1. **Simple moving average (SMA)-**
   It calculates the average of a fixed number of data points over a specified period, equally weighted.

2. **Weighted moving average (WMA)-** Like SMA, but assigns different weights to each data point, typically giving more importance to recent data.

3. **Exponential moving average (EMA)-** It gives more weight to recent data points, using an exponential decay factor to calculate the average, making it more responsive to recent changes in the data compared to SMA or WMA.

International Journal of Humanities Social Science and Management (IJHSSM)
Volume 4, Issue 2, Mar.-Apr., 2024, pp: 1115-1122 www.ijhssm.org

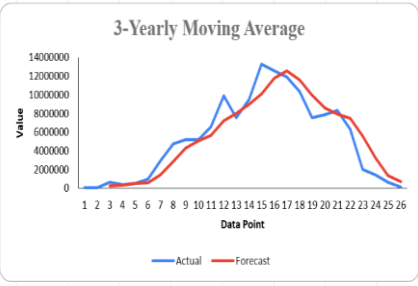
Window size choice affects sensitivity to data changes: larger windows provide smoother averages but may lag sudden changes, while smaller windows offer more responsiveness but may be noisy. It finds applications in finance, economics, engineering, and other fields for forecasting and trend analysis.

For our project, we took into consideration 3 Year Moving Average, 4 Year Moving Average and 5 Year Moving Average. For in depth analysis, we have considered a 5 yearly approach to make our understanding easier. Given below are our findings and analysis:

- **3 YEARLY MOVING AVERAGE**

3 YEAR

Year	Mileage	Moving Average
1989	31300	#N/A
1990	41250	#N/A
1992	609183	227244.3333
1994	380299	343577.3333
1995	483138	490873.3333
1996	935929	599788.6667
1997	2892443	1437170
1998	4716167	2848179.667
1999	5193396	4267335.333
2000	5220978	5043513.667
2001	6551324	5655232.667
2002	9880992	7217764.667
2003	7533128	7988481.333
2004	9488619	8967579.667
2005	1.3E+07	10096196.33
2006	1.3E+07	11764904
2007	1.2E+07	12573215.33
2008	1E+07	11593804.33
2009	7543958	9928706.667
2010	7834896	8569154.333
2011	8336059	7904971
2012	6332487	7501147.333
2013	2018928	5562491.333
2014	1424763	3258726
2015	609631	1351107.333
2016	114000	716131.3333

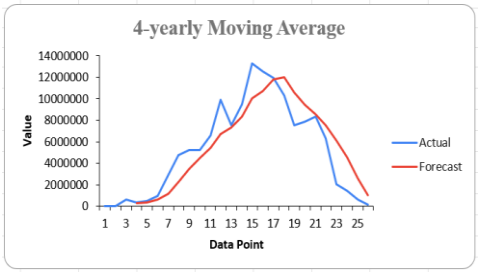


- **4 YEARLY MOVING AVERAGE**

4 YEAR

Year	Mileage	Moving Average
1989	31300	#N/A
1990	41250	#N/A
1992	609183	#N/A
1994	380299	265508
1995	483138	378467.5
1996	935929	602137.25
1997	2892443	1172952.25
1998	4716167	2256919.25
1999	5193396	3434483.75
2000	5220978	4505746
2001	6551324	5420466.25
2002	9880992	6711672.5
2003	7533128	7296605.5
2004	9488619	8363515.75
2005	1.3E+07	10042395.25
2006	1.3E+07	10706960
2007	1.2E+07	11802066.25
2008	1E+07	12012063.75
2009	7543958	10581342.75
2010	7834896	9405254
2011	8336059	8510880.5
2012	6332487	7511850
2013	2018928	6130592.5
2014	1424763	4528059.25
2015	609631	2596452.25
2016	114000	1041830.5

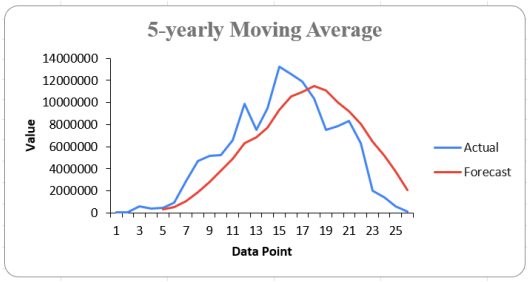
- **5 YEARLY MOVING AVERAGE**



International Journal of Humanities Social Science and Management (IJHSSM)
Volume 4, Issue 2, Mar.-Apr., 2024, pp: 1115-1122 www.ijhssm.org

5 YEAR

Year	Mileage	Moving Average
1989	31300	#N/A
1990	41250	#N/A
1992	609183	#N/A
1994	380299	#N/A
1995	483138	309034
1996	935929	489959.8
1997	2892443	1060198.4
1998	4716167	1881595.2
1999	5193396	2844214.6
2000	5220978	3791782.6
2001	6551324	4914861.6
2002	9880992	6312571.4
2003	7533128	6875963.6
2004	9488619	7735008.2
2005	1.3E+07	9344181
2006	1.3E+07	10541766.4
2007	1.2E+07	10948278.6
2008	1E+07	11507374.8
2009	7543958	11118442.6
2010	7834896	10032053.4
2011	8336059	9191415
2012	6332487	8075201.8
2013	2018928	6413265.6
2014	1424763	5189426.6
2015	609631	3744373.6
2016	114000	2099961.8



In the graphs given above, the blue shows the actual mileage, and the red line is the predicted mileage calculated using the moving average technique. Here we use the three yearly moving average to get an idea.

In earlier years, the proximity of the line to the axis suggests a notable trend: a conservative approach towards car usage prevailed. This conservatism stemmed from the limited production of automobiles, with cars being regarded more as luxuries than everyday commodities. However, as time advanced, the landscape of automobile manufacturing underwent significant transformations. The evolution of car production techniques led to mass production, marking a pivotal shift in societal attitudes towards car ownership and usage.

Over time, the trendline depicting car usage exhibits a noticeable upward trajectory, mirroring the broader societal shift towards embracing cars as essential assets rather than mere symbols of luxury. This shift was propelled by several factors, including advancements in manufacturing processes, increased accessibility to automobiles, and evolving consumer preferences. As cars became more accessible and affordable, they transcended their status as luxury items, permeating into the fabric of everyday life for a growing segment of the population.

The correlation between the upward trend of car usage and the parallel growth of the trendline underscores the symbiotic relationship between technological advancement and societal behaviour. The proliferation of automobiles not only reshaped transportation dynamics but also catalysed profound changes in urban development, economic structures, and cultural norms. The steady rise in car ownership reflects not only the democratization of mobility but also the profound impact of industrial innovation on shaping modern lifestyles.

Moreover, the expanding network of roads and highways, coupled with improvements in automotive technology, further facilitated the integration of cars into daily life. This integration sparked a cascade of secondary effects, ranging from the decentralization of urban centres to the emergence of suburban sprawl. Consequently, the once-distant trendline now intersects with the axis at increasingly higher points, symbolizing the deep-rooted transformation in societal attitudes towards car ownership and usage.

The narrative of the trendline's ascent parallels the broader narrative of industrial progress and societal evolution. What was once a symbol of privilege and exclusivity has now become a ubiquitous presence, underscoring the transformative power of technological innovation on reshaping human behaviour and the built environment. As we continue to chart the trajectory of car usage, it serves as a poignant reminder of the intricate interplay between innovation, culture, and the fabric of everyday life.

VI. REGRESSION ANALYSIS

Regression analysis is a statistical technique utilized to explore the association between multiple variables. Its application spans across diverse domains including economics, finance, biology, psychology, and sociology. The

International Journal of Humanities Social Science and Management (IJHSSM)
Volume 4, Issue 2, Mar.-Apr., 2024, pp: 1115-1122 www.ijhssm.org

fundamental objective of regression analysis is to comprehend the extent and character of the correlation between a dependent variable and one or more independent variables.

Linear regression is the prevailing form of regression analysis, wherein the relationship between variables is depicted through a straight line. The formula for a basic linear regression model, with y as the dependent variable and x as the independent variable, can be expressed as follows:

y = a + bx
Where:

- Y: dependent variable
- X: independent variable
- a: intercept.
- b: slope of the coefficient

likewise, the formula for a simple linear regression model with x on y can be represented as:
x = a + by

Regression analysis is valuable as it helps researchers to investigate and understand the relationship between variables in a quantitative manner.
It allows researchers to do:

- **Relationship Examination:** It helps researchers determine the nature and strength of the relationship between two or more variables. This is crucial for understanding how changes in one variable might affect
- **Prediction:** Regression analysis can be used to predict the value of one variable based on the values of other variables. This predictive capability is valuable in various fields such as finance, economics, and marketing for forecasting purposes.

- **Control:** By including multiple independent variables in the analysis, researchers can control for potential confounding factors and isolate the effect of a specific variable on the dependent variable.

In our project we have taken the two variables as time-Years (independent variable) and milage (dependent variable).
The following analysis is conducted to highlight the relationship and movement between our independent and dependent variables.

YEAR	Y ESTIMATE
2017	4680785.905
2018	4489717.782
2019	4298649.659
2020	4107581.535
2021	3916513.412
2022	3725445.289
2023	3534377.165
2024	3343309.042
2025	3152240.919

The regression equation has helped us to plot a straight line that best fits the data points (Actual sum of mileage). This trend line slopes downwards, which suggests that the sum of mileage has decreased over the years.

For example, according to the table, the mileage record in 2016 was 4871854.029 and the estimated mileage record in 2025 is 3152240.919. This is an estimated decrease of 1719613.11 over nine years. A falling trend can be indicative of the falling demand for used cars, even if the mileage is increasing.

Through our daily data, we could find the summation of mileage for various years as shown above. We can use the equation y=a+bx where x is time(years), and y is denoting the mileage of that year. We can find two constants a and b by finding the intercept (a) and slope (b) of the equation y on x.

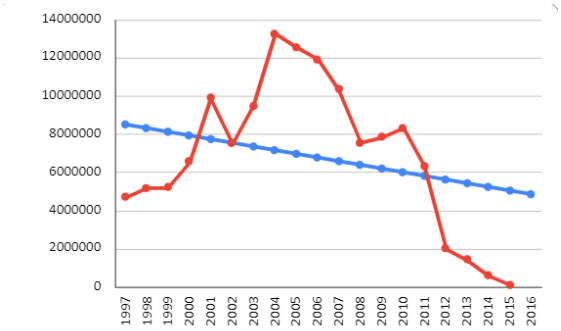
	Coefficients
Intercept	6687001
X Variable 1	-191068

Here we can see we got:
a=6687001 and b=-191068
Now we use these constants in our equation of y on x to derive the trend of the mileage throughout the given time, and hence this trend can be used to predict the mileage for future time.

International Journal of Humanities Social Science and Management (IJHSSM)
Volume 4, Issue 2, Mar.-Apr., 2024, pp: 1115-1122 www.ijhssm.org

Year(x)	Mileage(y)	X= year - 2006.5	Y=A+BX
1997	2892443	-9.5	8502148.371
1998	4716167	-8.5	8311080.248
1999	5193396	-7.5	8120012.125
2000	5220978	-6.5	7928944.002
2001	6551324	-5.5	7737875.878
2002	9880992	-4.5	7546807.755
2003	7533128	-3.5	7355739.632
2004	9488619	-2.5	7164671.508
2005	13266842	-1.5	6973603.385
2006	12539251	-0.5	6782535.262
2007	11913553	0.5	6591467.138
2008	10328609	1.5	6400399.015
2009	7543958	2.5	6209330.892
2010	7834896	3.5	6018262.768
2011	8336059	4.5	5827194.645
2012	6332487	5.5	5636126.522
2013	2018928	6.5	5445058.398
2014	1424763	7.5	5253990.275
2015	609631	8.5	5062922.152
2016	114000	9.5	4871854.029

We can plot the mileage and estimated Y values on a graph to create a graphical representation of the regression analysis. The blue line represents the trendline of mileage of used cars and the red line represents the actual mileage in their respective years.



VII. INDEX NUMBERS

Index numbers serve as a measurement tool for tracking changes in various variables over time or across distinct categories in the field of statistics. These numbers provide valuable insights into the relative change in a variable, typically expressed in percentage form.

Index numbers are widely utilized in economic analysis, where they help gauge the economic status of a particular region by comparing variables relative to a base period. They offer a means to assess changes in factors that cannot be directly measured or estimated. This makes them essential for studying the effects of such factors over time.

Furthermore, index numbers are valuable for tracking changes, facilitating comparisons between diverse groups or periods, and providing insights into trends and patterns in data. By representing relative changes, they enable analysts to understand how variables evolve and to make informed decisions based on the observed trends.

For our index number analysis, we have opted for the chain-based indices as our measure as our data lacks a structured breakdown of yearly prices for used cars, making it the only and most suitable measure for our purposes.

Chain-based indices are statistical measures used to track changes in a variable over time by taking a series of link relatives and "chaining" them together, allowing for a continuous comparison. It is calculated as (Current Year Link Relative* Previous Year Chain Based Index/100)
Link relatives are ratios that express the relationship between values in consecutive periods.
Here, **Link relative** expresses the change in mileage and engine displacement in a specific year relative to the previous year. It is calculated as:

$$\frac{\textbf{Current Year Price}}{\textbf{Previous Year Price}} \times \mathbf{100}$$

They are used to construct the index for each period, allowing for a continuous comparison of changes over time.

In this case, 1989 is taken as the base year, which is assigned a chain-based index number and link relative of 100.
The link relative shows the annual increase in the sum of mileages as well as engine displacement.

International Journal of Humanities Social Science and Management (IJHSSM)
Volume 4, Issue 2, Mar.-Apr., 2024, pp: 1115-1122 www.ijhssm.org

MILEAGE

MILEAGE			
Year	Price	Link Relative	Chain Based Index Number
1989	31300	100	100
1990	41250	131.7891374	131.7891374
1992	609183	1476.807273	1946.271565
1994	380299	62.42771056	1215.01278
1995	483138	127.0416173	1543.571885
1996	935929	193.7187719	2990.188498
1997	2892443	309.0451306	9241.031949
1998	4716167	163.0513376	15067.6262
1999	5193396	110.1190013	16592.31949
2000	5220978	100.5310976	16680.44089
2001	6551324	125.4807816	20930.7476
2002	9880992	150.8243525	31568.66454
2003	7533128	76.23858009	24067.5016
2004	9488619	125.9585527	30315.07668
2005	13266842	139.8184709	42386.07668
2006	12539251	94.51571821	40061.50479
2007	11913553	95.01008473	38062.46965
2008	10328609	86.69629455	32998.7508
2009	7543958	73.03943832	24102.10224
2010	7834896	103.8565697	25031.61661
2011	8336059	106.3965495	26632.77636
2012	6332487	75.96499737	20231.58786
2013	2018928	31.88207098	6450.249201
2014	1424819	70.57304669	4552.13738
2015	609631	42.78655745	1947.702875
2016	114000	18.69983646	364.2172524

- The year 1992 recorded the highest link relative. The link relative in 1992 was 1476.80, which shows the sum of mileage in 1992 was 1376.8% higher than that in 1989.
- The year 2016 recorded the lowest link relative. The link relative in 2016 was 18.69, which shows the sum of mileage in 2016 fell by 81.31% from that in 1989.
- Except for the notable spike observed in 1992, the link relatives for the sum of mileages exhibited minimal variation over time.
- The recession period of the early 1990s made people conscious of their expenses and thus they opted for fuel-efficient transportation. Thus, the higher sum of mileage in 1992.
- The year 2005 recorded the highest chain-based index number. The chain-based index number in 2005 was 42386.07, which shows the sum of mileage in 2005 was 42186.07% higher than that in 1989.
- The year 1990 recorded the lowest chain-based index number. The link relative in 1990 was 138.79, which shows the sum of mileage in 1990 increased by only 31.79% than that in 1989.
- The data for chain-based index numbers is highly varied. The sum of mileage increases till 2005 and then starts decreasing.
- The reason for the eventual decline can be the changing consumer preferences. Consumers started demanding safety and technological features, which were not found in used cars, even with high mileage.

ENGINE DISPLACEMENT

ENGINE DISPLACEMENT			
1989	1289	100	100
1990	1300	100.8533747	100.8533747
1992	5669	436.0769231	439.7982933
1994	2706	47.73328629	209.9301784
1995	2849	105.2845528	221.0240497
1996	8559	300.4212004	664.0031032
1997	30731	359.0489543	2384.096199
1998	38849	126.4163223	3013.886734
1999	49688	127.9003321	3854.77114
2000	47890	96.38142006	3715.283165
2001	70621	147.465024	5478.743212
2002	96020	135.9652228	7449.185415
2003	79321	82.60883149	6153.685027
2004	106180	133.8611465	8237.393328
2005	152885	143.9866265	11860.74476
2006	145549	95.20162213	11291.62141
2007	147388	101.263492	11434.29015
2008	139397	94.57825603	10814.35221
2009	107065	76.80581361	8306.051202
2010	118915	111.0680428	9225.368503
2011	110695	93.08749947	8587.664856
2012	97697	88.25782556	7579.286268
2013	38078	38.97560826	2954.072925
2014	73486	192.9880771	5701.008534
2015	115622	157.3388128	8969.899147
2016	1197	1.035270104	92.86268425

- The year 1992 recorded the highest link relative. The link relative in 1992 was 436.07, which shows the sum of engine displacement in 1992 was 336.07% higher than that in 1989.
- The year 2016 recorded the lowest link relative. The link relative in 2016 was 1.03, which shows the sum of engine displacement in 2016 fell by 98.97% from that in 1989.
- The sum of engine displacement exhibited a fluctuating pattern over time, characterized by both increases and decreases.

- These fluctuations can be due to the fluctuations in the demand and supply for certain types of cars. For example, if there is a demand for larger vehicles such as trucks or SUVs, there may be an increase in engine displacement.
- The year 2005 recorded the highest chain-based index number. The chain-based index number in 2005 was 143.98, which shows the sum of engine displacement in 2005 was 43.99% higher than that in 1989.
- The year 2016 recorded the lowest chain-based index number. The link relative in 2016 was 92.86, which shows the sum of engine displacement in 1990 decreased by 7.14% from that in 1989.
- The data for chain-based index numbers is highly varied.

VIII. LIMITATIONS

This project makes use of a variety of statistical methods to analyse the secondary data that was collected by the group. It includes the following:

1. Moving Averages
2. Least Square Method
3. Index Numbers

A **Moving Average** is a tool to analyse a given set of data points by creating an average of the observations in the given period. It creates stability in data and removes the fluctuations.

Moving averages are slow to react to the current movements in data as they are based on historical data. They can miss identifying key turning points in the data set. Also, it misses key fluctuation points.

Moving averages fail to predict future behaviour of the data as they are made from historical figures.

Least Square Method is used to determine and analyse a regression equation and establishing trend lines based on the same.

Regression Analysis helps in establishing a relationship between a dependent variable and one or more independent variables. However, it has some limitations such as:

This method assumes that the errors in the data are random/normally distributed and presumes that the mean is 0. If these are not fulfilled, then the result can be misleading.

The method is also sensitive to extreme data points that can deviate the results. Also, least square method compares the relation of two variable and may be unable to capture the whole data efficiently.

Index numbers are used as a statistical measure to track changes in data points of individual and grouped goods or services over a period. They express this change as a percentage relative to a chosen base period. In simpler terms, they tell you how much more expensive (or cheaper) a basket of goods is compared to a specific point in the past.

One of the limitations of the index numbers is that people may substitute a good for a cheaper one and this effect is not reflected on the index.

Another drawback of this method is that a particular base year may have a much greater impact on the interpretation of index compared to other years. It also overlooks the changes in quality and only considers the changes in price of the product/service.

IX. CONCLUSION

In conclusion, the utilization of time series analysis and index numbers in examining the data of used cars has yielded invaluable insights into the dynamics of the consumers of the market. Through the meticulous examination of trends over time and the construction of meaningful indices, we have been able to discern patterns, fluctuations, and overall market behaviour with greater clarity.

Time series analysis has provided us with a comprehensive understanding of how numerous factors influence the pricing and demand for used cars over time. By observing trends, seasonality, and potential cyclical patterns, we can make informed predictions and decisions regarding future market movements.

Additionally, the application of index numbers has enabled us to condense complex data into concise and meaningful measures, facilitating comparisons and assessments of relative changes in prices, quantities, and other pertinent variables across different time periods and regions.

Overall, the combination of time series analysis and index numbers has empowered us to make evidence-based conclusions about the used car market, assisting stakeholders in making informed decisions regarding pricing strategies, inventory management, and investment opportunities. Moving forward, these analytical techniques will continue to be instrumental in navigating the ever-evolving landscape of the automotive industry.

REFERENCES:

[1]. https://www.kaggle.com/datasets/mirosval/personal-cars-classifieds