IMPACT OF COVID 19 ON CAR SALES IN INDIA

CONTENTS

INTRODUCTION

Car manufacturing is one of the major industries in India, it contributes about 7 percent of the country's GDP with an annual turnover of 7.5 lakh crore and export of Rs.3.5 lakh crore. The automotive and mobility industries have certainly been among the hardest hit during the COVID-19 pandemic.Sales had been dropping for 12 to 15 months when the outbreak stalled production and overall economic activity.

Customers aspiring to buy entry-level cars or those looking for an upgrade to such models are postponing decisions as the COVID-19 pandemic has hit their income sentiment significantly. At the same time, sales of premium segment cars are expected to go up on account of resilient income of affluent.

In India the leading car manufacturer is Maruthi Suzuki India Limited, a subsidiary of Suzuki Motor Corporation,Japan. The eight Maruthi Suzuki models contribute over 80 percentage of the top 10 selling models and above 60 percent of total car sales in India is contributed by Maruthi. The COVID-19 Pandemic has cased negative growth for the industry. Some of the immediate short-term concerns for the industry are pandemic-related uncertainties and health our people,global shortage of semiconductors,rising commodity prices,upcoming fuel-efficiency and Bharat-Stage VI(BS-VI) phase 2 regulations,short of shipping containers and import restrictions.

OBJECTIVE

The main objective of this study is to analyse the car sales data during COVID-19.The data we collected is from May-2020 to Feb-2022. We are doing forecasting the next 5 months data using R software. Also we analyse the customer buying preference over car during COVID-19, here we are considering the first 10 cars in the ranking order of number of unit sold in May-2022. We use Multiple Linear Regression model to find out the significant factors that affect the sales.

TOOLS AND SOFTWARE

To achieve the objective of our study we need various statistical tools and software like Excel,RStudio etc.These are

- MS Excel
- R Software
- Line Graph
- Bar Chart
- Pie Chart

TERMINOLOGY

- Time Series

  A time series is a data set that tracks a sample over time. In particular, a time series allows one to see what factors influence certain variables from period to period.

- Stationary Process

  A Stationary process in time series is that the data does not depend at which the series is observed. In general a stationary time series will have no predictable pattern in long term. It does not mean that the series does not change over time, just that the way it changes does not itself change over time.

- Autocorrelation

  Autocorrelation refers to the degree of correlation of the same variables between two successive time intervals. It measures how the lagged version of the value of a variable is related to the original version of it in a time series. Autocorrelation is also known as serial correlation.

  The formula of Theoretical Autocorrelation is

$$\rho_k = \frac{E(X_t - \mu)(X_{t-k} - \mu)}{\sigma_x^2}$$

  The formula for sample Autocorrelation is

$$r_k = \frac{\sum_{t=1}^{n-k}(X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2}$$

- Partial Autocorrelation

  A partial autocorrelation is a summary of the relationship between an observation in a time series with observations at prior time steps with the relationships of intervening observations removed.Partial Autocorrelation for lag 2 is

$$\Phi_{22} = PACF(X_t, X_{t-2})$$

$$=\text{Cor}(X_t, X_{t-2}/Xt-1)$$

$$= \frac{Cov(X_t, X_{t-k}/Xt-1)}{\sqrt{Var(X_t/X_{t-1})}\sqrt{Var(X_{t-2}/X_{t-1})}}$$

- Autoregressive Intergrated Moving Average(ARIMA

  An ARIMA model is a generalization of a simple Autoregressive Moving Average(ARMA) model.ARIMA model is generally denoted as ARIMA(p,q,d) where p,d,q are defined as follows

- p: the lag order or the number of time lag of autoregressive model AR(p)
- d: degree of differencing or the number of times the data have had subtracted with past value
- q: the order of moving average model MA(q)

- Forecasting

  Forecasting is a technique that use historical data as inputs to make informed estimates that are predictive in determining the direction of future trends. It is used by utilities for analyzing their data and predict whether their sales,share or profit trend is up or down.

- Multiple Linear Regression Analysis

  Multiple Linear Regression or simply known as multiple regression is a statistical technique that used several explanatory variables to predict the outcome of response variable.

ANALYSIS

- DATA SITUATION

Category : Sub-segment	Models	June 2020			April-June 2020			April'19 - March'20
		2020	2019	% Change	2020-21	2019-20	% Change	
A: Mini	Alto, S-Presso[2]	10,458	18,733	-44.2%	12,453	57,893	-78.5%	247,776
A: Compact	WagonR, Swift, Celerio, Ignis, Baleno, Dzire, Tour S	26,696	62,897	-57.6%	32,958	205,178	-83.9%	787,610
Mini + Compact Segment		**37,154**	**81,630**	**-54.5%**	**45,411**	**263,071**	**-82.7%**	**1,035,386**
A: Mid-Size	Ciaz	553	2,322	-76.2%	745	8,703	-91.4%	25,258
Total A: Passenger Cars		**37,707**	**83,952**	**-55.1%**	**46,156**	**271,774**	**-83.0%**	**1,060,644**
B: Utility vehicles	S-Cross, Vitara Brezza, XL6[2], Ertiga	9,764	17,797	-45.1%	13,400	58,984	-77.3%	235,298
C: Vans	Omni, Eeco	3,803	9,265	-59.0%	5,420	32,659	-83.4%	118,404
Total Domestic Passenger Vehicle Sales		**51,274**	**111,014**	**-53.8%**	**64,976**	**363,417**	**-82.1%**	**1,414,346**
Light Commercial Vehicles	Super Carry	1,026	2,017	-49.1%	1,189	6,568	-81.9%	21,778
Total Domestic Sales (PV+LCV)		**52,300**	**113,031**	**-53.7%**	**66,165**	**369,985**	**-82.1%**	**1,436,124**
Sales to other OEM: A: Compact		839	1,830	-54.2%	862	4,496	-80.8%	25,002
Total Domestic Sales (Domestic + OEM)[1]		**53,139**	**114,861**	**-53.7%**	**67,027**	**374,481**	**-82.1%**	**1,461,126**
Total Export Sales		**4,289**	**9,847**	**-56.4%**	**9,572**	**28,113**	**-66.0%**	**102,171**
Total Sales (Total Domestic + Export)[1]		**57,428**	**124,708**	**-54.0%**	**76,599**	**402,594**	**-81.0%**	**1,563,297**

Clarifications:

Table: Monthly Sales Report by Maruthi

Manufacturer	Bodystyle	Model	Sales	Mileage/L	seat capacity
Maruti Suzuki	Hatchback	Wagon R	16814	23.56	5
Tata	Hatchback	Nexon	14614	21.19	5
Maruti Suzuki	Hatchback	Swift	14133	23.2	5
Maruti Suzuki	Hatchback	Baleno	13970	26	5
Maruti Suzuki	Hatchback	Alto	12933	22.05	5
Maruti Suzuki	MUV	Ertiga	12226	20.3	7
Maruti Suzuki	Sedan	Dzire	11603	23.26	5
Hyundai	SUV	Creta	10973	16.8	5
Maruti Suzuki	Van	Eeco	10482	16.2	7
Maruti Suzuki	SUV	Brezza	10312	17.03	5

Sunroof	First service cost	Ground Clearance	Waiting Period(weeks)	wheel size(Inch)
No	1249	165	5	14
Yes	2590	209	12	16
No	1574	163	6	15
No	1331	170	6	16
No	2671	160	7	12
No	1899	185	8	15
No	1625	170	12	15
Yes	2800	190	26	16
No	2817	160	7	13
No	3220	200	10	16

Table:

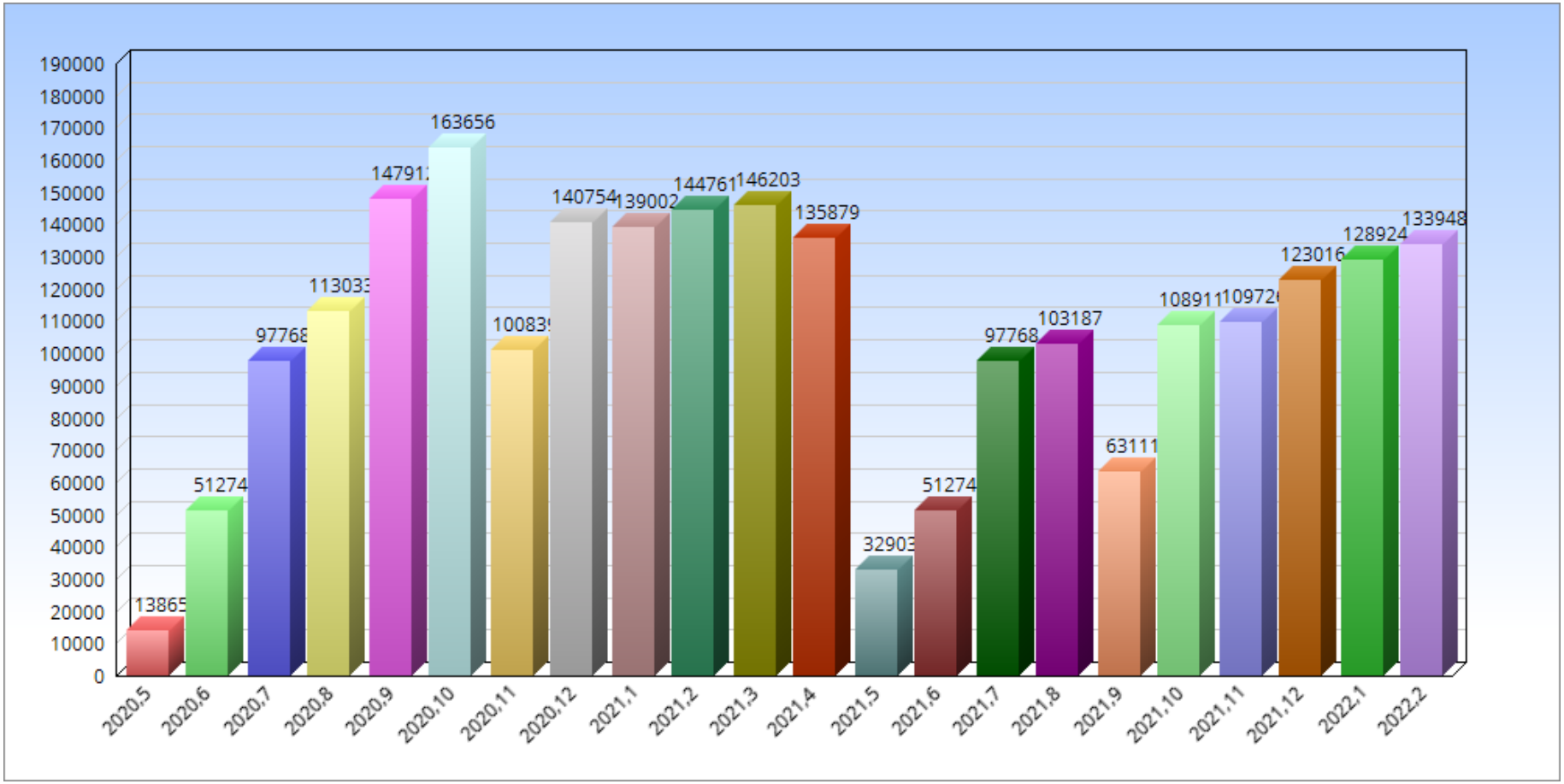

Figure: Car Sales of Months

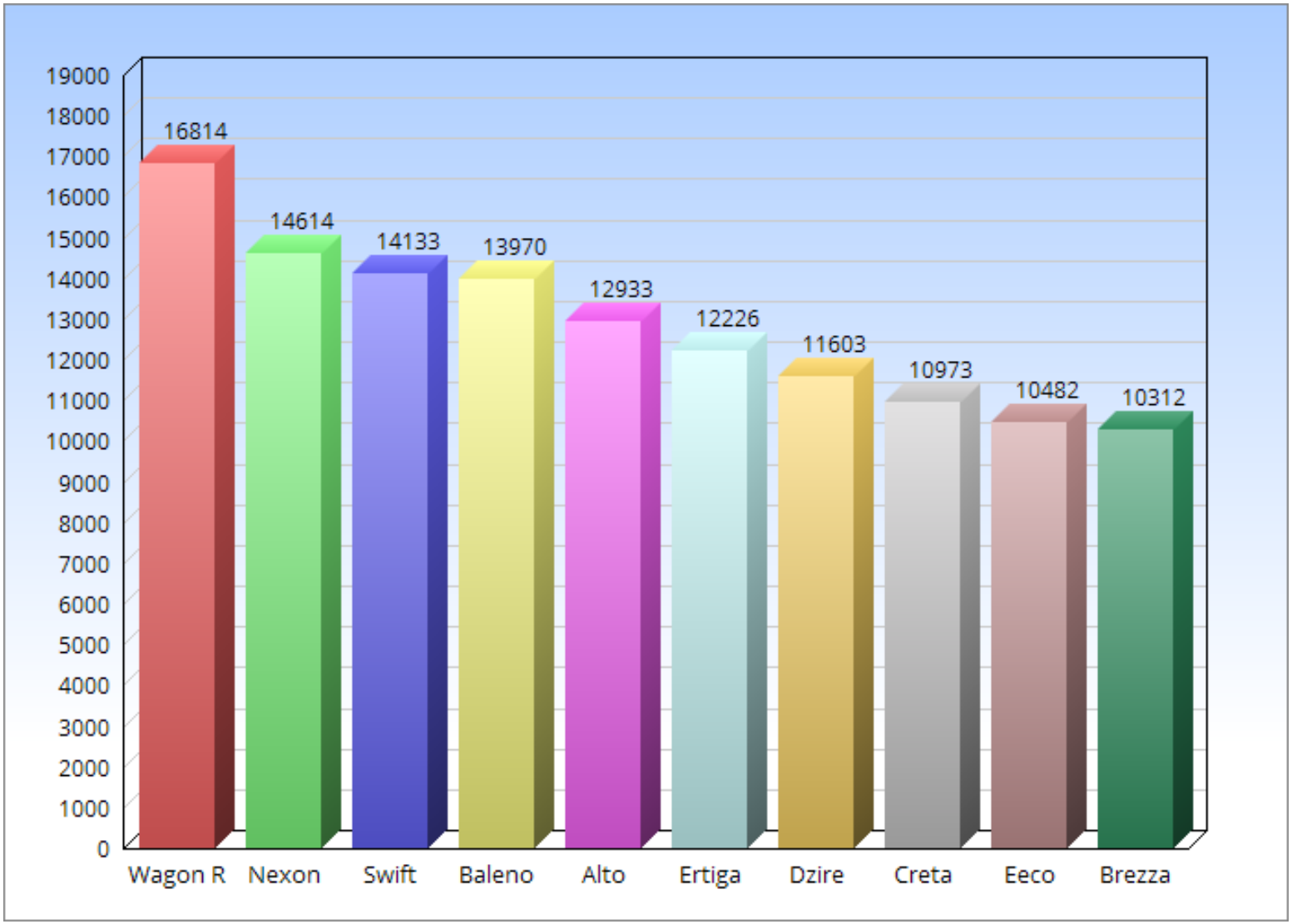

Figure: Car Sales in May-2022

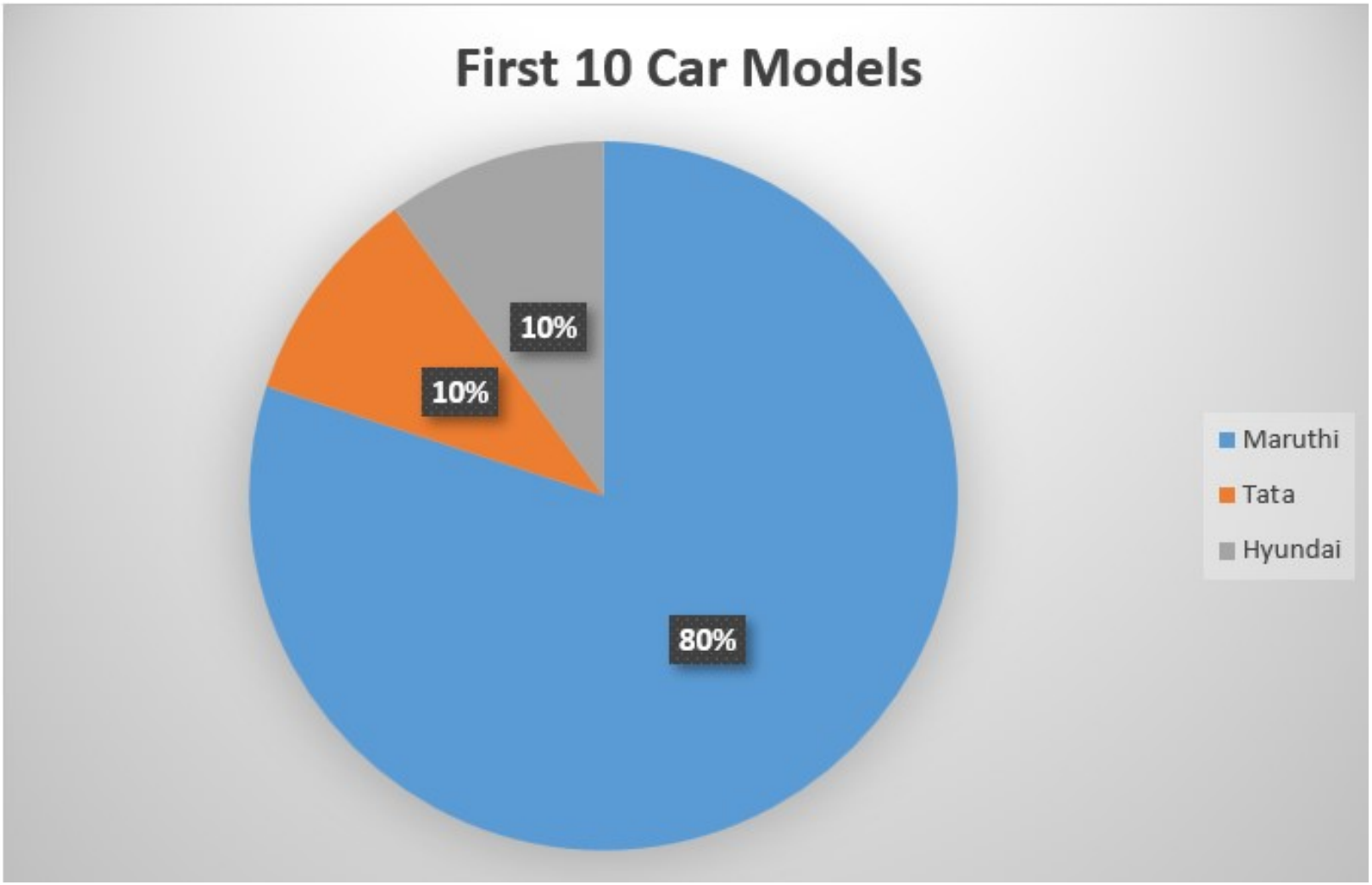

Figure: Pie Chart of Manufacturer

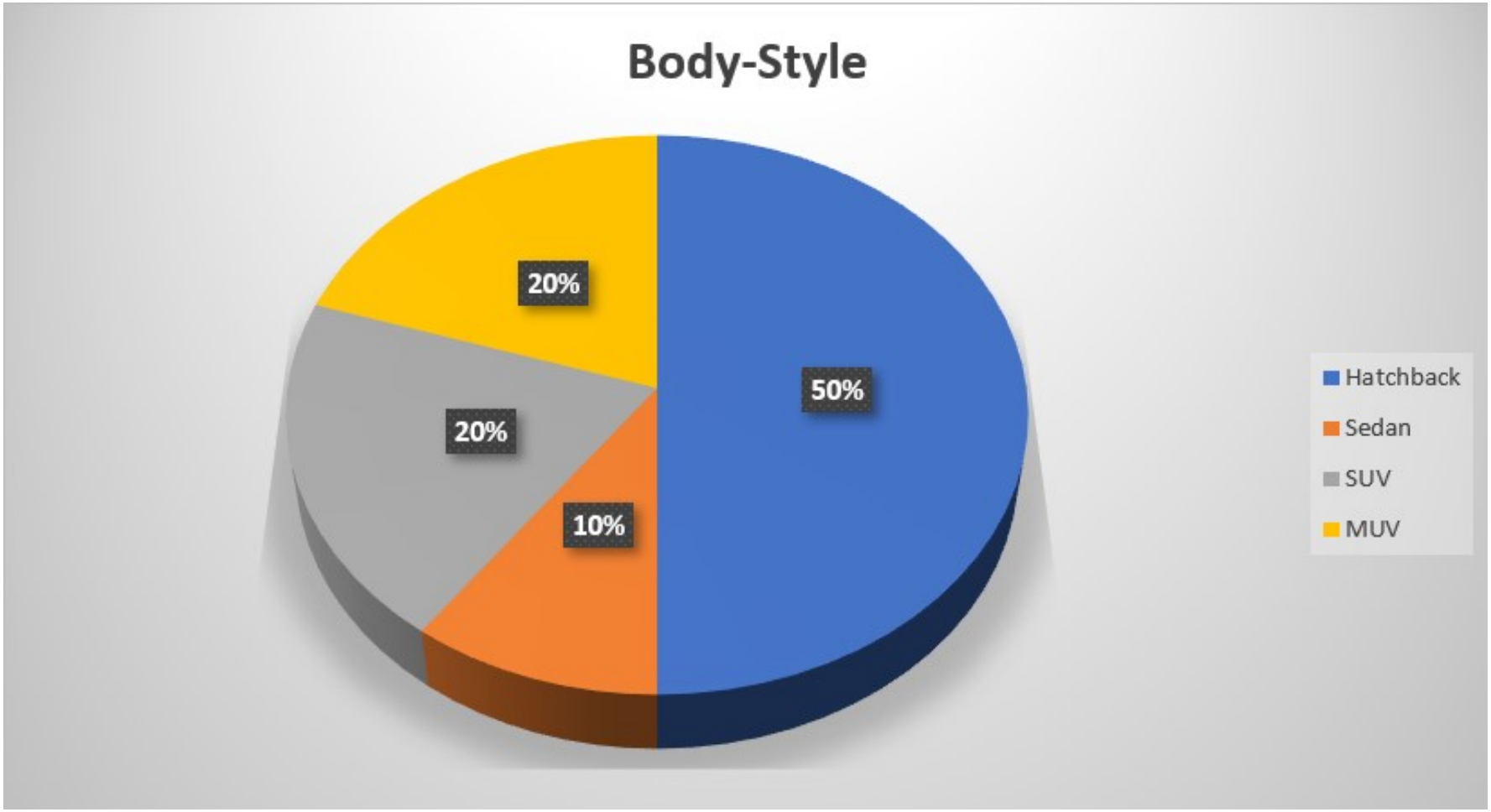

Figure: Pie Chart of Body Style

- Data Processing

  First, we want to import the data into our R software from excel using the r code given below(fig:9). Later we plot the time series data in our r software and is given below(fig:10)

  > data=read.csv(file.choose(),header=T)

  > data

  > ts.plot(Data$Monthly.Sales,xlab="months")

  >acf(data$Monthly.Sales,lag.max=30)

  >pacf(data$Monthly.Sales,lag.max=30)

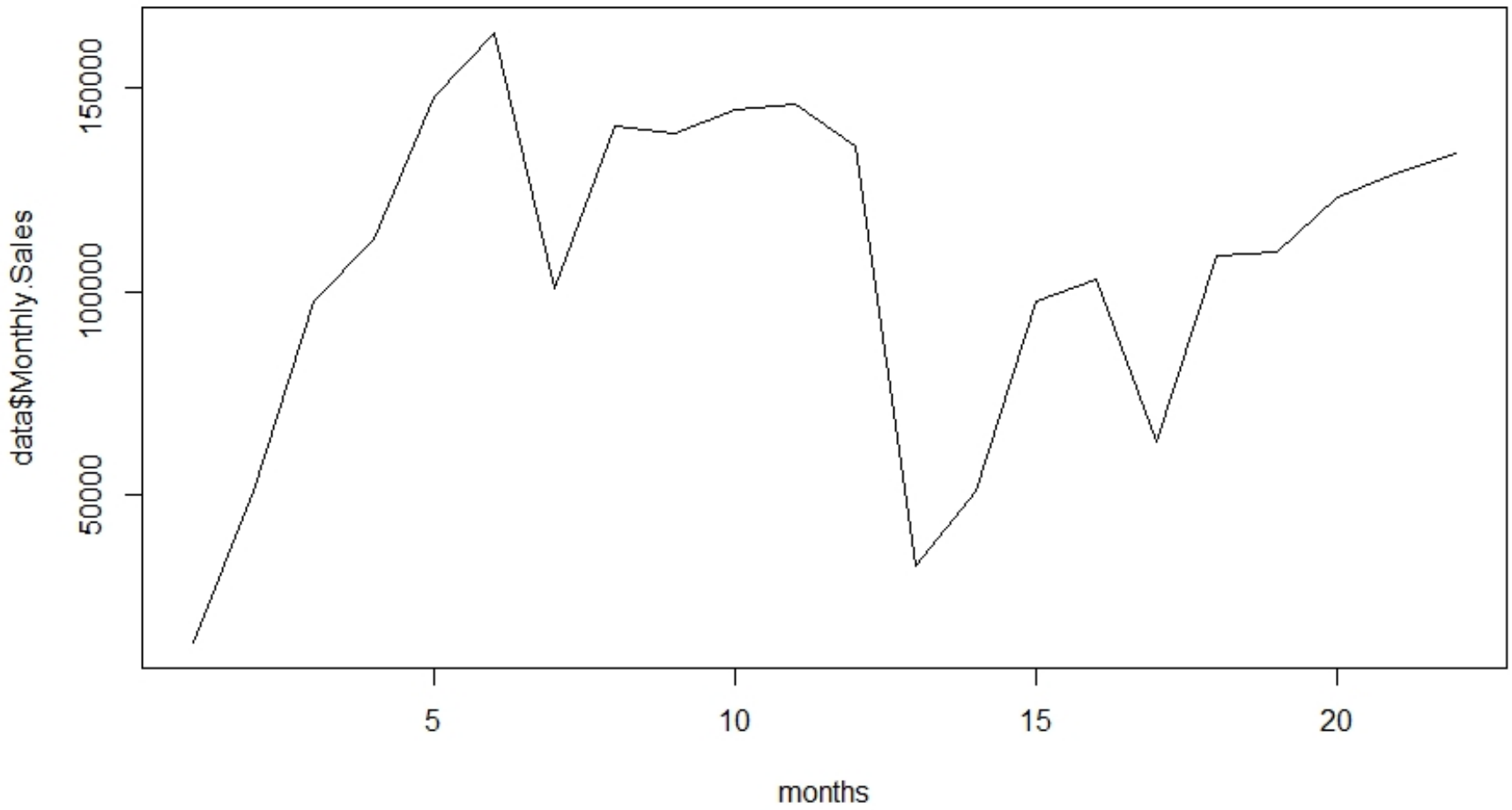

Figure: Time Series Plot

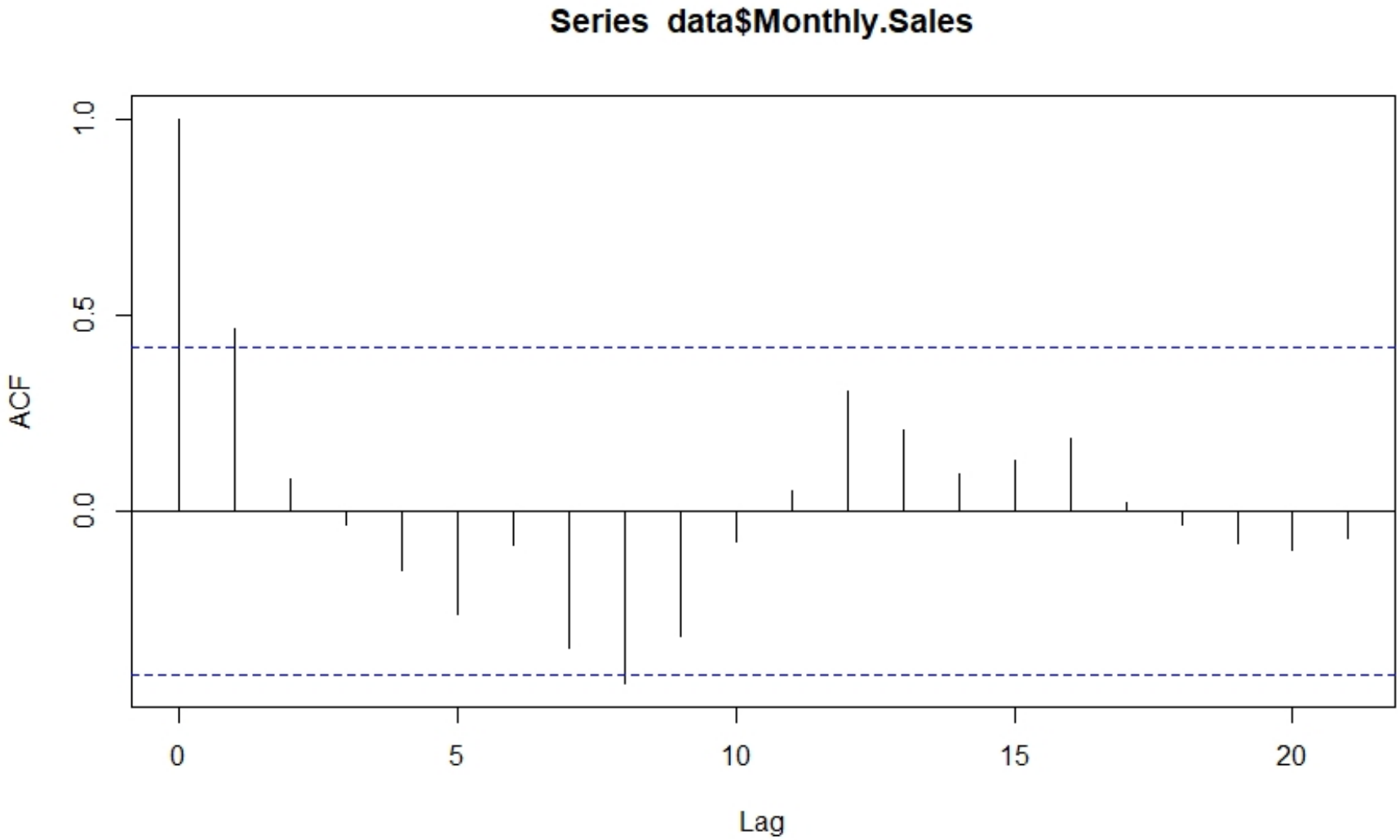

Figure: Autocorrelation graph of the sequence

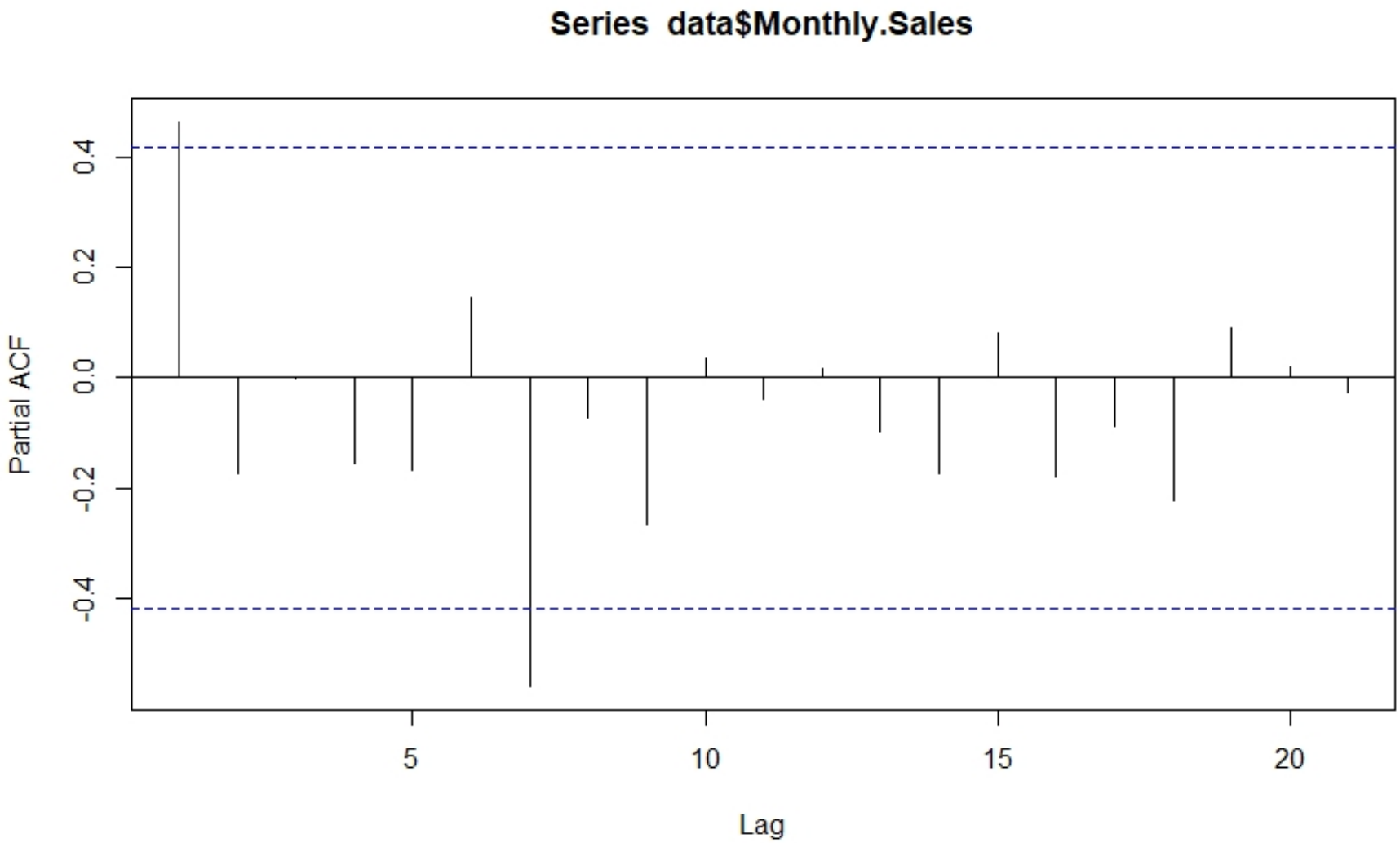

Figure: Partial ACF graph of the sequence

- INFERENCE

  We plot the ACF and PACF of the respective data, we observe that the ACF and PACF graph falls to zero, so we assume that the given data is stationary and make more evidence for our statement.

- Autocorrelation Test

  or stationary series of data, the ACF autocorrelation graph immediately tend to zero. The two dashed lines in the autocorrelation graph represent the confidence bounds, which are the upper and lower bounds of the autocorrelation coefficients. The R code for plotting ACF is given below

  > acf(Data$Monthly.Sales,lag.max=30)

- Kwiatkowski-Phillips-Schmidt-Shin (**KPSS**) Test is a statistical test for testing whether a data is stationary or not. We define the null and alternate hypothesis as

$$H_0\text{:Series is stationary or has no unit root}$$

$$H_1\text{:Series is non stationary or has a unit root}$$

We fail to reject null hypothesis if the p values is greater than significant level(0.05). The key difference of KPSS test from tests like ADF is that the null hypothesis is the given series is stationary whereas ADF test is vice-versa.

- KPSS Test in R Software

The R code for KPSS test is
>kpss.test(data$Monthly.Sales)

```
> kpss.test(data$Monthly.Sales)

        KPSS Test for Level Stationarity

data:  data$Monthly.Sales
KPSS Level = 0.10907, Truncation lag parameter = 2, p-value = 0.1

Warning message:
In kpss.test(data$Monthly.Sales) : p-value greater than printed p-value
```

Figure: KPSS Test in R

The p-value is greater than 0.05 so we fail to reject the null hypothesis. So let us conclude that the given data is stationarity

ARIMA MODELING ANALYSIS

- Identifying the Order

  Since we have obtained ACF and PACF of the data. In ACF we have two significant values and we have only one significant value in PCF.Since we haven't done differencing in our data, we assume three models.First we assume three models and check their Akaike and Bayesian Information Criterion(AIC & BIC)

- AIC estimates the quality of each model, when we represent a data while modeling some information about the data may be lost. AIC estimates the relative amount lost during this modeling.

- BIC:while fitting models to increase the likelihood we add more parameters which may fail to fit the model and predict adequately,BIC checks this criterion.
- ARIMA Modeling in R

  As mentioned above first we execute auto arima function in R. Before executing auto arima function we must convert our data into time series data in R. There are several methods for converting our data into a time series data in R. R code for converting the data into time series is

  >tsdata=ts(data[,2],start=c(2020,5), end = c(2022,2), frequency = 12)

```
> tsdata=ts(data[,2],start=c(2020,5), end = c(2022,2), frequency = 12)
> tsdata
        Jan    Feb    Mar    Apr    May    Jun    Jul    Aug    Sep    Oct    Nov    Dec
2020                             13865  51274  97768 113033 147912 163656 100839 140754
2021 139002 144761 146203 135879  32903  51274  97768 103187  63111 108911 109726 123016
2022 128924 133948
```

Figure: data converted to time series

Now we call the three ARIMA model

- ARIMA(1,0,0)
- ARIMA(1,0,1)
- ARIMA(0,0,2)

R code for calling ARIMA model and to find the AIC and BIC values

>arima1=arima(data$Monthly.Sales,order=c(1,0,1))

>AIC(arima1)

>BIC(arima1)

```
> arima1=arima(data$Monthly.Sales,order=c(1,0,0))
> arima2=arima(data$Monthly.Sales,order=c(1,0,2))
> arima3=arima(data$Monthly.Sales,order=c(0,0,2))
> AIC(arima1)
[1] 527.2266
> BIC(arima1)
[1] 530.4997
> AIC(arima2)
[1] 525.2343
> BIC(arima2)
[1] 530.6895
> AIC(arima3)
[1] 528.4122
> BIC(arima3)
[1] 532.7764
```

Figure: AIC and BIC of ARIMA

We find the BIC values of ARIMA(1,0,0) is less than the other two models whereas the AIC value of ARIMA(1,0,2) is less than the above model, therefore for more confirmation we do the auto arima function.

```
> auto.arima(tsdata,max.order = c(3,0,3),stationary = T,trace=T,ic='aicc')

 ARIMA(2,0,2)            with non-zero mean : Inf
 ARIMA(0,0,0)            with non-zero mean : 533.1143
 ARIMA(1,0,0)            with non-zero mean : 528.5599
 ARIMA(0,0,1)            with non-zero mean : 529.1407
 ARIMA(0,0,0)            with zero mean     : 576.9288
 ARIMA(2,0,0)            with non-zero mean : 530.8806
 ARIMA(1,0,1)            with non-zero mean : 530.9114
 ARIMA(2,0,1)            with non-zero mean : 534.2804
 ARIMA(1,0,0)            with zero mean     : 530.7332

 Best model: ARIMA(1,0,0)            with non-zero mean

Series: tsdata
ARIMA(1,0,0) with non-zero mean

Coefficients:
         ar1       mean
      0.5974  102805.72
s.e.  0.1959   16850.92

sigma^2 = 1.229e+09:  log likelihood = -260.61
AIC=527.23   AICc=528.56   BIC=530.5
```

Figure: Auto ARIMA in R

Here the auto arima function checked the models up to the order (3,0,3) and found out that the best order for your ARIMA is (1,0,0) that is Auto Regressive model of order 1. Now we want to forecast and predict using this model.

FORECASTING

Now we have identified our ARIMA model with evidence. Our next procedure is to forecast the sales data in R. In our forecasting procedure, we predict the next 5 months data using R. The R code for forcasting is

```
>forecast=forecast:::forecast.Arima(arima,h=5,level=c(80,90))
>forecast
>plot(forecast)
```

We have to forecast the 5 month sales data, since the data we choose is from 5-2020 to 2-2022 we can use Mean absolute percentage error (MAPE) to measure forecast accuracy. MAPE is the sum of the individual absolute errors divided by the demand, that is the average of the percentage error. We have the 3 months data (March,April,May) after 2-2020 so we compare these original data and our predicted data. A good MAPE value score is as follows

- $<$10 percent:Very Good
- 10-20 percent:Good
- 20-50 percent:Ok
- $>$50 percent:Bad

```
> forecast=forecast:::forecast.Arima(arima,h=5,level=c(80,90))
> forecast
   Point Forecast    Lo 80    Hi 80    Lo 90    Hi 90
23       121409.7 78572.55 164246.8 66428.82 176390.5
24       113919.4 64020.74 163818.1 49875.16 177963.7
25       109444.9 57256.96 161632.8 42462.41 176427.4
26       106771.9 53790.92 159752.8 38771.56 174772.2
27       105175.0 51913.95 158436.1 36815.17 173534.9
> plot(forecast)
```

Figure: forecast of Sales in R

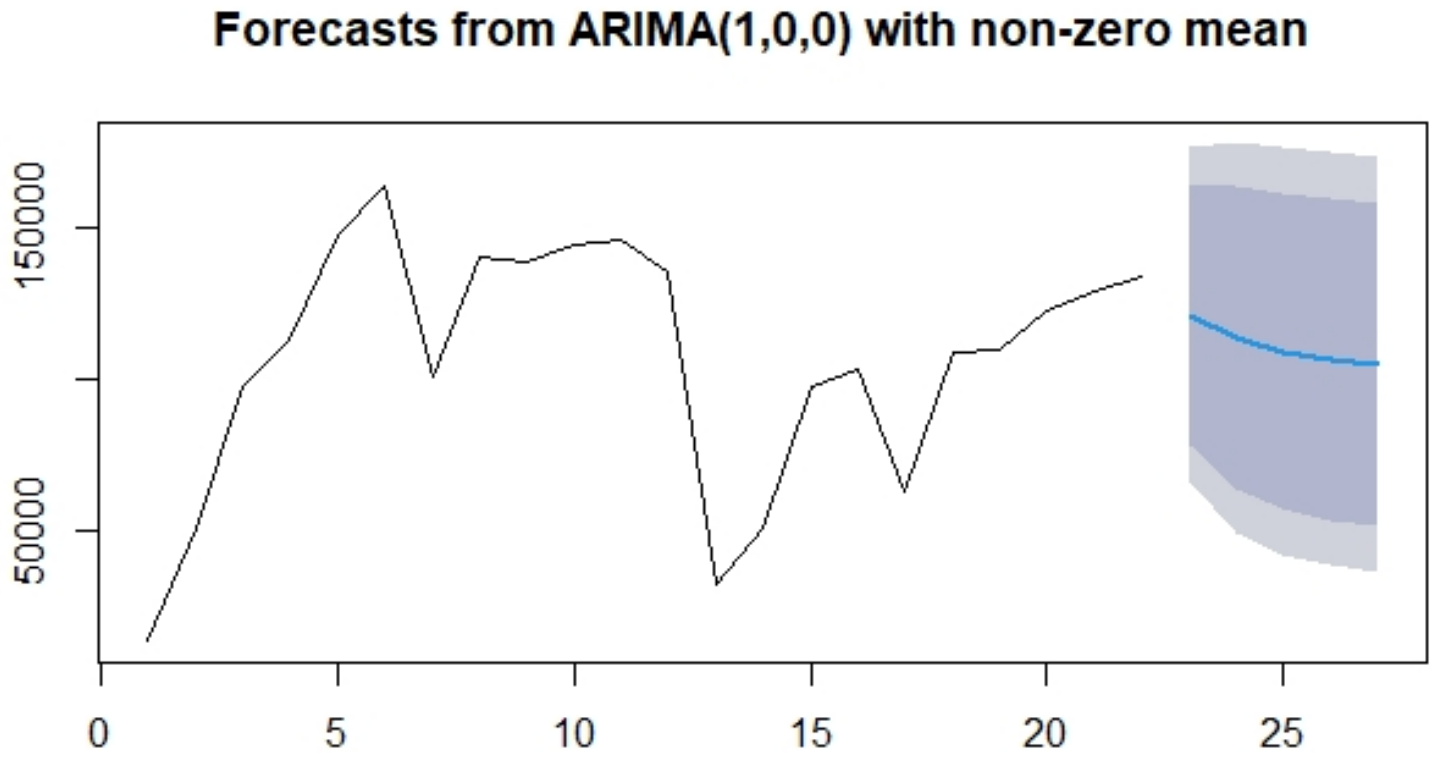

Figure: Plot of forecast in R

```
> orginal_data=c(133816,121995,124474)
> predicted_data=c(121409,113919,109444)
> MAPE(orginal_data,predicted_data)
[1] 0.1034716
```

Figure: MAPE test in R

- Inference

  From Figure the predictions can be clearly seen,we draw the primitive and predictive graphics in r using plot command. Also we get the value of MAPE Measure as 10.34 percent which indicate our forecasting is good.

Multiple Linear Regression Analysis

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...$$

Where Y is the response variable,$X_1, X_2, ...$ are the regressor variable, β_0 is the Y-Intercept and $\beta_1, \beta_2, ...$ are the regression coefficient

We state the null and alternate hypothesis as

$$H_0 : \beta_1 = \beta_2 = ... = 0$$

H_1:At least one $\beta_j \neq 0$

- Multiple Linear Regression in R

  After importing our data into R software, we check the correlations of the factors and choose the factors in which sales have more correlations.Later we fit the model using lm() function in R and call. If the p-value is $>$0.05 the we fail to reject null hypothesis or else we accept H_1.

  The R code for fitting the model is

  >model=lm(data$Sales data$Mileage.L+data$Bodystyle.1+ data$seat.capacity+ data$First.service.cost.min.10k.km.)

  >summary(model)

```
> model=lm(data$Sales~data$Mileage.L+data$Bodystyle.1+data$seat.capacity+data$First.service
.cost.min.10k.km.)
> summary(model)

Call:
lm(formula = data$Sales ~ data$Mileage.L + data$Bodystyle.1 +
    data$seat.capacity + data$First.service.cost.min.10k.km.)

Residuals:
      1       2       3       4       5       6       7       8       9      10
1072.83  742.05 -1026.88 -110.18 -225.80  973.43 -904.04  -27.09 -973.43  479.11

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      28817.278  11011.318   2.617   0.0473 *
data$Mileage.L                    -598.353    373.224  -1.603   0.1698
data$Bodystyle.1                 -2491.918    793.978  -3.139   0.0257 *
data$seat.capacity                1314.960    760.773   1.728   0.1445
data$First.service.cost.min.10k.km.   -2.451      1.189  -2.063   0.0941 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1073 on 5 degrees of freedom
Multiple R-squared:  0.8531,    Adjusted R-squared:  0.7356
F-statistic:  7.26 on 4 and 5 DF,  p-value: 0.0259
```

Figure: Model fitting in R

- Inference From above test p-value is less than 0.05 so we reject null hypothesis, that is we accept that some of the regression coefficients are significant.In our Model the Intercept and Body Style is significant

  The regression coefficients are

  β_0=28817.278

  β_1=-2491.918

  where β_0 is the y-intercept and β_1 is Body style.

  Therefore our model is given by

$$Y = 28817.278 + -2491.918X_1$$

CONCLUSION

After analysis,the best model is ARIMA(1,0,0) that is the AR-1 model. Since our data is stationary we need not do differencing. This ARIMA(1,0,0) model is used for predict the short-term trend of domestic passenger segment cars of Maruthi-Suzuki.When MAPE is measured we got the forecasting is good. In Multiple Linear Regression modeling we confirmed that some of the factors mentioned which depends the sales are significant.that is the Body style.The factors are selected according their correlations.

Forecasting shows that there is a possibility of a decrease in sales. To overcome this the company must do new marketing strategies like giving exciting offers.Company must alsDuring COVID-19 restriction the company must give an option for online purchases and provide home delivery even in remote places.From monthly car sales data, it is seen that customer prefer Hatchback over other models. So to increase the number of sales the company must do more research and development on giving more features in hatchbacks. The Manufacturer must give more authorized customization options for their cars in both exterior and interior. Availing Engine modifications through an authorized service center attracts youth to buy that respective cars.