**Business Analytics using R Case Study**

**Introduction –**

In this case study, we delve into a dataset containing information on car purchases by 1000 customers. This data offers valuable insights into key factors influencing consumer behavior in the car market.

The dataset encompasses details such as User ID, Gender, Age, Annual Salary, and a binary Purchased variable indicating whether a car was purchased or not. The customer base represents a diverse range of individuals, with genders equally distributed (male and female), ages ranging from 18 to 63 years old, and annual salaries varying between Rs. 15,000 and Rs. 152,500.

Our primary objective is to analyze this car sales data and identify the crucial factors that shape consumer decisions. By uncovering these factors, we aim to:

* Gain a deeper understanding of the target market for car purchases.
* Identify potential trends and correlations between various customer attributes and purchase decisions.
* Develop actionable strategies to increase sales by targeting the right audience and addressing their needs.

This analysis will provide valuable insights for car manufacturers, dealerships, and marketing teams, enabling them to tailor their offerings and strategies to attract and convert potential customers more effectively.

**Objective –**

To analyze car sales data for identifying key factors influencing consumer behavior and recommending strategies to increase sales.

**Importing Dataset –**

df<-read.csv('D:\\ishita\\college\_r\\car\_data.csv',header=T)

> head(df)

User.ID Gender Age AnnualSalary Purchased

1 385 Male 35 20000 0

2 681 Male 40 43500 0

3 353 Male 49 74000 0

4 895 Male 40 107500 1

5 661 Male 25 79000 0

**Data Cleaning –**

**Creating categorical variables –**

> df$Purchased\_Status<-factor(df$Purchased, levels=c(0,1), labels=c('Not Purchased', 'Purchased'))

> df$Age\_Group<-ifelse(df$Age>=18 & df$Age<=35, "Young", ifelse(df$Age>=36 & df$Age<=44, "Middle", "Senior"))

> df$Salary\_Group<-ifelse(df$AnnualSalary<30000, "Low", ifelse(df$AnnualSalary>=30000 & df$AnnualSalary<75000, "Middle", "High"))

> head(df)

User.ID Gender Age AnnualSalary Purchased Purchased\_Status Age\_Group Salary\_Group

1 385 Male 35 20000 0 Not Purchased Young Low

2 681 Male 40 43500 0 Not Purchased Middle Middle

3 353 Male 49 74000 0 Not Purchased Senior Middle

4 895 Male 40 107500 1 Purchased Middle High

5 661 Male 25 79000 0 Not Purchased Young High

6 846 Female 47 33500 1 Purchased Senior Middle

**Checking and Removing outliers in age and salary –**

> #checking outliers for salary

> salary <- df$AnnualSalary

> Q1 <- quantile(salary, 0.25)

> Q3 <- quantile(salary, 0.75)

> IQR <- Q3 - Q1

> lower\_bound <- Q1 - 1.5 \* IQR

> upper\_bound <- Q3 + 1.5 \* IQR

> outliers <- nrow(salary < lower\_bound | salary > upper\_bound)

> outliers

NULL

> #checking outliers for age

> age <- df$Age

> q1 <- quantile(age, 0.25)

> q3 <- quantile(age, 0.75)

> iqr <- q3 - q1

> lb <- q1 - 1.5 \* iqr

> ub <- q3 + 1.5 \* iqr

> Outliers <- nrow(age < lb | age > ub)

> Outliers

NULL

* *Since there are no outliers in both the variables - age and salary, hence we are moving forward with the current dataset.*

**Analysis & Interpretation –**

**Checking Normality of salary & age distribution for deciding the selection of parametric and non-parametric tests –**

> #checking normality for age

> v=mean(df$Age)

> w=sd(df$Age)

> e=v+w

> e

[1] 50.81307

> f=v-w

> f

[1] 29.39893

> g=nrow(df[df$Age>=f & df$Age<=e,])

> g

[1] 636

> h=(g/nrow(df))\*100

> h

[1] 63.6

* *Criteria 1 is not satisfied as area of the curve is* 63.6% which is less than 68.3%. This implies that the data is not normally distributed.

> #checking normality for salary

> a=mean(salary)

> b=sd(salary)

> c=a+b

> c

[1] 107177.3

> d=a-b

> d

[1] 38200.66

> e=nrow(df[salary>=d & df$Age<=c,])

> e

[1] 816

> f=(e/nrow(df))\*100

> f

[1] 81.6

* *Criteria 1 is satisfied as area of the curve is* 81.6% which is more than 68.3%.

> i=a+2\*b

> j=a-2\*b

> i

[1] 141665.7

> j

[1] 3712.316

> k=nrow(df[salary>=j & salary<=i,])

> k

[1] 954

> l=(k/nrow(df))\*100

> l

[1] 95.4

* *Criteria 2 is satisfied as area of the curve is* 95.4% which is more than 95%.

> p=a+3\*b

> q=a-3\*b

> p

[1] 176154

> q

[1] -30776.03

> r=nrow(df[df$AnnualSalary>=q & df$AnnualSalary<=p,])

> r

[1] 1000

> s=(r/nrow(df))\*100

> s

[1] 100

* *Criteria 3 is satisfied as area of the curve is* 100% which is more than 99.7%.
* *Thus, Annual salary has a normal distribution.*
* Conclusion thereby is that we can perform parametric test for salary and non parametric for age.

**Does gender-specific preferences effects the consumer purchasing decisions?**

*H0 – There is no significant difference in car purchases between males and females.*

*H1 – There is a significant difference in car purchases between males and females.*

> obs\_freq <- table(df$Gender, df$Purchased\_Status)

> obs\_freq

Not Purchased Purchased

Female 297 219

Male 301 183

> row\_sums<-rowSums(obs\_freq)

> col\_sums<-colSums(obs\_freq)

> total<-sum(obs\_freq)

> exp\_freq<-row\_sums %o% col\_sums/total

> exp\_freq

Not Purchased Purchased

Female 308.568 207.432

Male 289.432 194.568

> chi\_sq<-sum((obs\_freq-exp\_freq)^2/exp\_freq)

> chi\_sq

[1] 2.228919

> dof<-(nrow(obs\_freq)-1)\*(ncol(obs\_freq)-1)

> dof

[1] 1

*\*\* critical value at 5% significance level – 3.841*

* *Value of chi-sq =2.2289 which is less than 3.841. Thus, H0 is accepted.*

**Is age of the customer a prerequisite for the purchase of the car?**

*H0 – There is no significant difference in car purchases between different age groups.*

*H1 – There is a significant difference in car purchases between different age groups.*

> obs\_freq <- table(df$Age\_Group, df$Purchased\_Status)

> obs\_freq

Not Purchased Purchased

Middle 248 92

Senior 57 279

Young 293 31

> row\_sums<-rowSums(obs\_freq)

> col\_sums<-colSums(obs\_freq)

> total<-sum(obs\_freq)

> exp\_freq<-row\_sums %o% col\_sums/total

> exp\_freq

Not Purchased Purchased

Middle 203.320 136.680

Senior 200.928 135.072

Young 193.752 130.248

> chi\_sq<-sum((obs\_freq-exp\_freq)^2/exp\_freq)

> chi\_sq

[1] 407.3521

> dof<-(nrow(obs\_freq)-1)\*(ncol(obs\_freq)-1)

> dof

[1] 2

*\*\* critical value at 5% significance level – 5.991*

* *Value of chi-sq =* 407.3521 *which is more than 5.991. Thus, H1 is accepted.*

**Is salary of the customer a prerequisite for the customer to purchase car?**

*H0 – There is no significant difference in car purchases between different salary groups.*

*H1 – There is a significant difference in car purchases between different salary groups.*

> obs\_freq <- table(df$Salary\_Group, df$Purchased\_Status)

> obs\_freq

Not Purchased Purchased

High 186 258

Low 72 41

Middle 340 103

> row\_sums<-rowSums(obs\_freq)

> col\_sums<-colSums(obs\_freq)

> total<-sum(obs\_freq)

> exp\_freq<-row\_sums %o% col\_sums/total

> exp\_freq

Not Purchased Purchased

High 265.512 178.488

Low 67.574 45.426

Middle 264.914 178.086

> chi\_sq<-sum((obs\_freq-exp\_freq)^2/exp\_freq)

> chi\_sq

[1] 112.8933

> dof<-(nrow(obs\_freq)-1)\*(ncol(obs\_freq)-1)

> dof

[1] 2

*\*\* critical value at 5% significance level – 5.991*

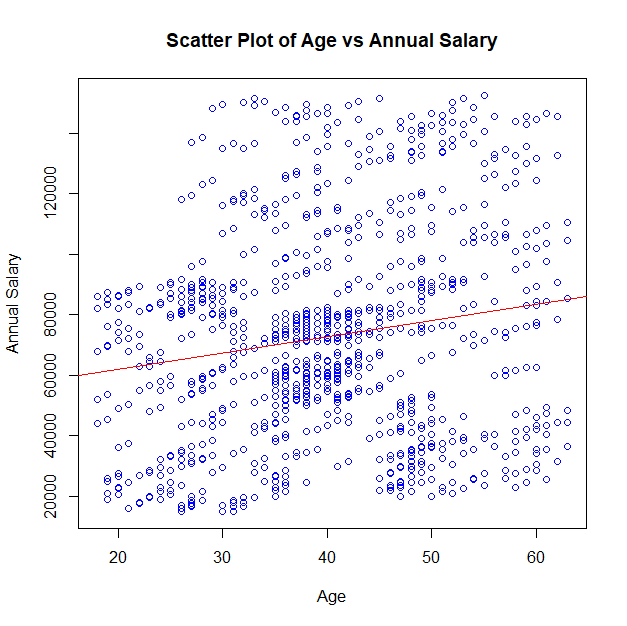
* *Value of chi-sq =* 112.8933 *which is more than 5.991. Thus, H1 is accepted.*

**How far is the age and the salary of the customer related?**

> plot(df$Age, df$AnnualSalary,main = "Scatter Plot of Age vs Annual Salary",xlab = "Age",ylab = "Annual Salary",col = 'blue')

> model <- lm(AnnualSalary ~ Age, data = df)

> abline(model, col = 'red')



* *The above scatter plot shows the direct relationship between age and annual salary which means that as age increases salary also increases.*

> summary\_model <- summary(model)

> r\_squared <- summary\_model$r.squared

> cat("R-squared:", r\_squared, "\n")

R-squared: 0.02757002

* *This value indicates that approximately 2.76% of the variation in salary can be explained by the variation in age. In other words, age is a very weak predictor of salary.*

> adjusted\_r\_squared <- summary\_model$adj.r.squared

> cat("Adjusted R-squared:", adjusted\_r\_squared, "\n")

Adjusted R-squared: 0.02659564

* *This value is similar to R-squared but adjusts for the number of predictors in the model. In this case, the adjusted R-squared is slightly lower, suggesting that the model's predictive power might be even weaker when considering the complexity of the model.*

> predicted <- predict(model)

> actual <- df$AnnualSalary

> mape <- mean(abs((actual - predicted) / actual)) \* 100

> cat("Mean Absolute Percentage Error (MAPE):", mape, "%\n")

Mean Absolute Percentage Error (MAPE): 55.71426 %

* *This metric measures the accuracy of the model's predictions. A lower MAPE indicates better accuracy. In this case, the MAPE of 55.71% suggests that, on average, the model's predictions are off by about 55.71%. This is a relatively high MAPE, indicating that the model's predictions are not very accurate.*

Based on these statistical measures, the model that uses age to predict salary is not a strong predictor. The model explains only a small portion of the variation in salary, and its predictions are significantly inaccurate.

> # **Pearson correlation** coefficient

> mean\_age <- sum(age) / length(age)

> mean\_salary <- sum(salary) / length(salary)

> sd\_age <- sqrt(sum((age - mean\_age)^2) / (length(age) - 1))

> sd\_salary <- sqrt(sum((salary - mean\_salary)^2) / (length(salary) - 1))

> covariance <- sum((age - mean\_age) \* (salary - mean\_salary)) / (length(age) - 1)

> correlation <- covariance / (sd\_age \* sd\_salary)

> print(correlation)

[1] 0.1660422

* *A Pearson correlation coefficient of 0.17 indicates a weak positive correlation between age and salary. This means that as age increases, there is a slight tendency for salary to also increase, but the relationship is not very strong.*

> m=matrix(c(1,.17,.17,1),ncol=2,nrow=2,byrow=TRUE)

> colnames(m) <- c("age", "salary")

> rownames(m) <- c("age", "salary")

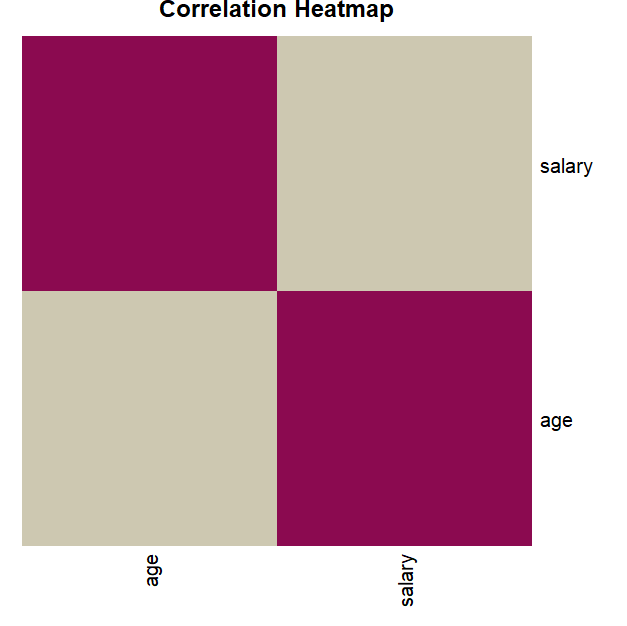
> m

age salary

age 1.00 0.17

salary 0.17 1.00

> heatmap(m,Colv=NA,Rowv=NA,col= colorRampPalette(c("deeppink4", "cornsilk3"))(100),main = "Correlation Heatmap",cexRow = 1.5,cexCol = 1.5)



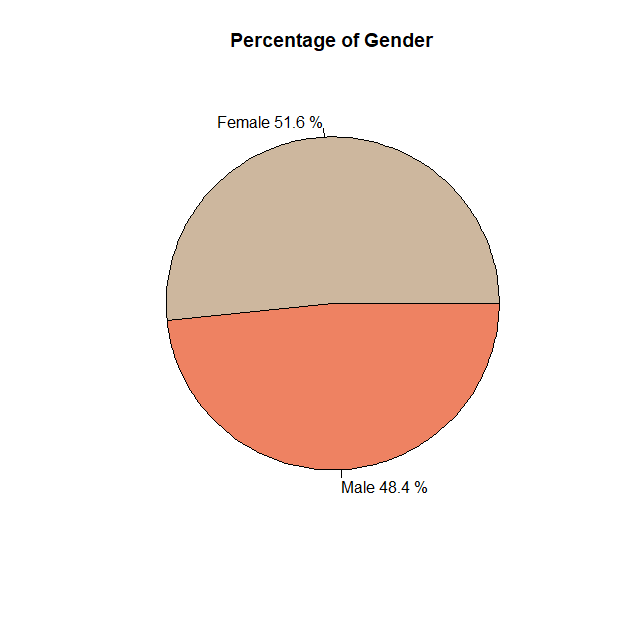
**Visual Analysis of Gender -**

> table(df$Gender)

Female Male

516 484

>pie(table(df$Gender),labels=paste(names(table(df$Gender)),(c(table(df$Gender))\*100)/1000,'%',sep=" "),main='Percentage of Gender',col=c("bisque3", "salmon2"))



* *The pie chart illustrates the gender distribution of the sample, revealing that 51.6% of the participants are female, while 48.4% are male, showcasing a near-equal balance between the two genders in the dataset.*

> table<-table(df$Purchased\_Status,df$Gender)

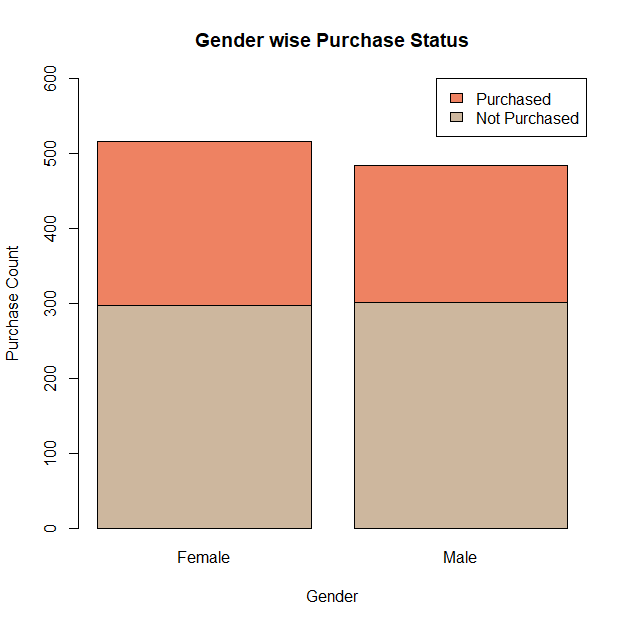
> table

Female Male

Not Purchased 297 301

Purchased 219 183

> barplot(table,col = c("bisque3", "salmon2"),legend.text = c("Not Purchased", "Purchased"),args.legend = list(x = "topright"),xlab = "Gender",ylab = "Purchase Count",main = "Gender wise Purchase Status", ylim=c(0,600))



* *The barplot illustrates that gender does not affect whether a customer will purchases the car or not.*

**Visual Analysis of Age-**

> table(df$Age\_Group)

Middle Senior Young

340 336 324

> table<-table(df$Purchased\_Status,df$Age\_Group)

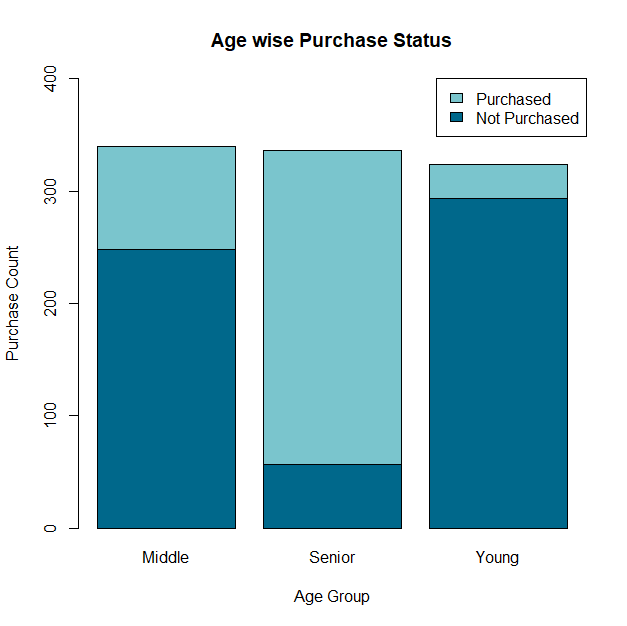
> table

Middle Senior Young

Not Purchased 248 57 293

Purchased 92 279 31

> barplot(table,col=c('deepskyblue4','cadetblue3','dodgerblue4'),legend.text = c("Not Purchased", "Purchased"),args.legend = list(x = "topright"),xlab = "Age Group",ylab = "Purchase Count",main = "Age wise Purchase Status", ylim=c(0,400))

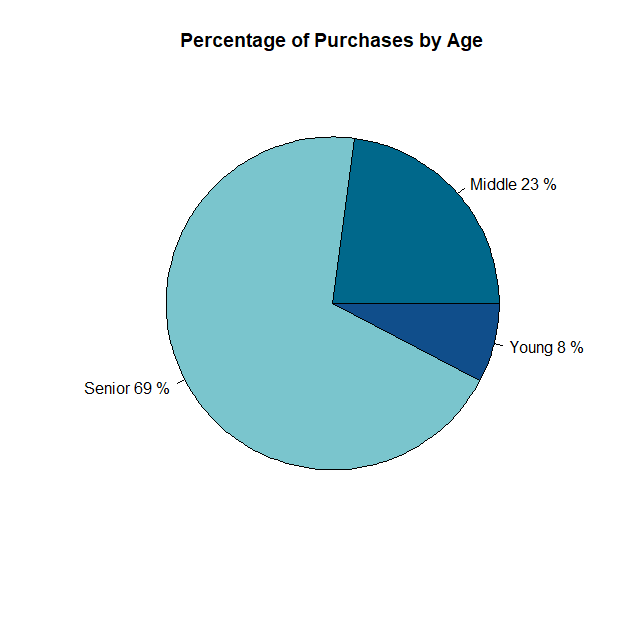


* *The barplot reveals that senior citizens constitute the largest segment of car buyers, with a total of 279 purchases*. *In contrast, young adults represent the smallest group, with only 31 purchases.*

> purchases <- table(df$Age\_Group[df$Purchased == 1])

> purchase\_percent <- prop.table(purchases) \* 100

> pie(purchase\_percent,labels = paste(names(purchase\_percent), round(purchase\_percent, 0), "%"),main = "Percentage of Purchases by Age",col=c('deepskyblue4','cadetblue3','dodgerblue4'))



* *The pie chart displays the distribution of car purchasers across different age groups, highlighting a significant preference for car ownership among senior citizens compared to middle and younger age groups.*

**Visual Analysis of Salary–**

> table(df$Salary\_Group)

High Low Middle

444 113 443

> table<-table(df$Purchased\_Status,df$Salary\_Group)

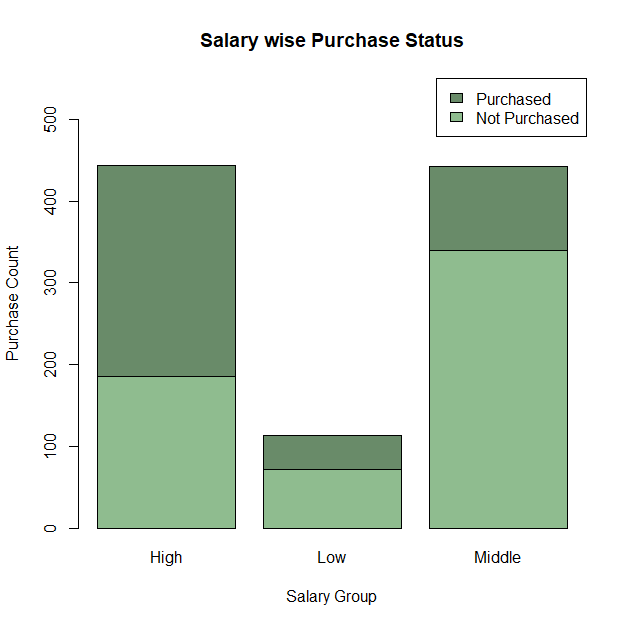
> table

High Low Middle

Not Purchased 186 72 340

Purchased 258 41 103

> barplot(table,col=c('darkseagreen','darkseagreen4','aquamarine4'),legend.text = c("Not Purchased", "Purchased"),args.legend = list(x = "topright"),xlab = "Salary Group",ylab = "Purchase Count",main = "Salary wise Purchase Status", ylim=c(0,550))

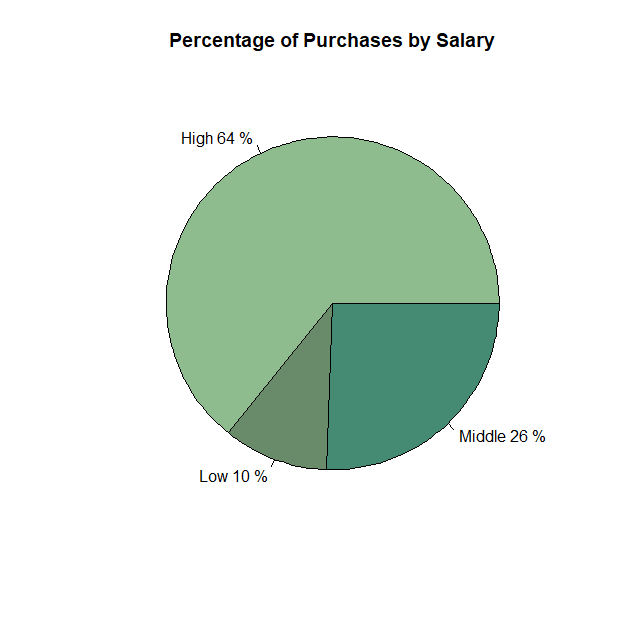


* *The barplot shows that individuals from the high salary group are the most prominent buyers, with a total of 258 purchases. Conversely, the low salary group represents the smallest segment, with only 41 purchases.*

> purchases <- table(df$Salary\_Group[df$Purchased == 1])

> purchase\_percent <- prop.table(purchases) \* 100

> pie(purchase\_percent,labels = paste(names(purchase\_percent), round(purchase\_percent, 0), "%"),main = "Percentage of Purchases by Salary",col=c('darkseagreen','darkseagreen4','aquamarine4'))



* *This pie chart illustrates the distribution of car purchasers based on salary groups, indicating that higher income levels are associated with increased car ownership.*

**Which variable among gender, salary and age do you think has the maximum impact on the purchasing decision of car by the customer?**

* **Salary** appears to have the maximum impact on the purchasing decision of a car.
* Salary:
  + Individuals in the high-salary group were the most likely to purchase cars.
  + Individuals in the low-salary group were the least likely to purchase cars.
  + This indicates a strong correlation between income level and car purchasing behavior.
* Age:
  + While senior citizens were the most likely to purchase cars, the impact of age on purchasing decisions seems less pronounced compared to salary.
  + Younger adults were less likely to purchase cars, but this could be due to various factors like financial constraints, lifestyle choices, and priorities.
* Gender:
  + Gender did not appear to have a significant impact on car purchase decisions, suggesting that both male and female customers are equally likely to purchase cars.

Therefore, while age and gender may influence purchasing decisions to some extent, salary appears to be the most significant factor in determining whether a customer will purchase a car.

**Insights –**

**1. Age and Salary:**

* There is a weak positive correlation between age and annual salary.
* Age is not a strong predictor of salary, as evidenced by the low R-squared and adjusted R-squared values.
* The model's predictions for salary based on age are significantly inaccurate, as indicated by the high MAPE.

**2. Gender and Purchase:**

* The dataset is nearly equally divided between male and female customers.
* Gender does not appear to have a significant impact on car purchase decisions.

**3. Age and Purchase:**

* Senior citizens (older age group) are the most likely to purchase cars.
* Younger adults are the least likely to purchase cars.

**4. Salary and Purchase:**

* Individuals in the high-salary group are the most likely to purchase cars.
* Individuals in the low-salary group are the least likely to purchase cars.

**Implications for Car Sales Strategy –**

Based on these findings, the following strategies can be considered to increase car sales:

**1. Target the Right Demographic:**

* **Prioritize Senior Citizens:** Focus marketing efforts on senior citizens, as they are the most likely to purchase cars.
* **Target High-Income Individuals:** Direct marketing campaigns towards individuals in the high-salary group.

**2. Understand the Impact of Age and Salary:**

* **Age-Based Marketing:** Tailor marketing messages and product offerings to different age groups.
* **Income-Based Segmentation:** Segment the market based on income levels and offer tailored financing options.

**3. Gender-Neutral Marketing:**

* Since gender does not significantly influence car purchase decisions, adopt gender-neutral marketing strategies.

**4. Product Positioning and Pricing:**

* **Value Proposition for Seniors:** Emphasize safety features, comfort, and ease of use in marketing to senior citizens.
* **Premium Offerings for High-Income Groups:** Offer high-end, luxury models and exclusive services to attract high-income customers.

By understanding these key factors and implementing targeted strategies, car manufacturers and dealerships can enhance their sales performance and better meet the needs of their customers.

**Conclusion –**

The analysis of the car purchase data revealed that salary was the most significant factor influencing purchasing decisions, with higher income individuals being more likely to purchase cars. Age and gender, while influential to some extent, had less impact. To optimize sales, car companies should target high-income individuals, tailor marketing strategies to different age groups, and adopt a gender-neutral approach.

*Submitted to – Mr. Saurabh Agarwal*

*Submitted by – Ishita Maheshwari*

*MBA II (B) [230230088]*