Maximizing Car Consumer Satisfaction Through Data-Driven Quality and Performance

hajiyousefi.kimiya@stud.hs-fresenius.de

HS-Fresenius: Data Science for Business

Contents

Rendered at 13 February, 2023

Word count: 6162

Abstract

Customer satisfaction is among the most crucial concerns in today's business world. Related terms such as customer loyalty, customer experience management, and customer success have also become integral aspects of modern organizations. It does not matter where you stand in the production and economic cycles; everything you do is always in line to increase customer satisfaction. In the meantime, car manufacturers in Iran are not exempted. Due to the monopoly of this industry in Iran, they are always trying to increase customer satisfaction or keep it at an acceptable level. This matter cannot be implemented in any way except with customer research and satisfaction research through an engineered vision that can reveal the unsaid for managers and make decision-making more reliable for them. This report aims to identify the most important and influential factors influencing consumer satisfaction. Hopefully, the examination and monitoring of customer satisfaction, which serves as an ostensible marker, is not just a marketing tool but a legitimate avenue to improve the quality of the automotive industry.

1 Introduction

In this study, I want to increase the customer satisfaction rate and solve a business issue in the car manufacturing industry in Iran. Iran's automotive industry is the second largest after the oil industry in Iran. Currently, there are two manufacturers, IKCO and Saipa, and they are the only car manufacturers controlling the market. This monopoly in Iran's automobile industry has created a unipolar market, resulting in a decrease in quality and an excessive increase in product prices. Due to needing the desired quality, the companies suffered a drop in sales. Naturally, one of the solutions they considered was to check customer satisfaction, find the points that lowered the quality and fix these defects. Since then, these businesses have been working to improve the quality of their products. Therefore, in this respect, I found some good literature that can help me understand the different aspects of the subject as much as possible. Gathering customer feedback on existing products and services will give businesses the insight to drive future decisions, resulting in a genuinely customer-oriented business. The customer satisfaction goals should be aimed at: improving customer loyalty, increasing customer satisfaction rates, increasing product advocacy, improving product usability, and driving successful cross-team collaboration (Bloemer and Lemmink 1992). The ultimate concern of most businesses in any industry in today's market-oriented business climate is how to satisfy customers. As a result, the mainstream academic literature about CS has been formed in recent years. Managers need to understand customer satisfaction (CS) dimensions, quantify them, and use these measurements. Due to its enormous effects on both the long-term performance of businesses and consumer purchasing habits, CS is crucial to measure. In the academic community, it is well known that regularly delivering high customer satisfaction (CS) is linked to greater customer loyalty and an improved reputation (Wangenheim and Bayon 2004), (Fornell 1992), (Anderson and Sullivan. 1993). The following is a commonly accepted definition of satisfaction: "Satisfaction is the consumer's fulfilling response." A product or service feature, or the product or service itself, "gave (or continues to give) a pleasurable level of consumption-related fulfilment, including levels of under- or over-fulfilment”(Oliver 1997) A product’s or service’s satisfaction is a construct that is dependent on use and experience (Oliver 1997). This definition is astounding. First, the “consumer” is the main subject rather than the “customer.” Traditionally, a customer pays for a product or service but may not be the consumer. The consumer (direct user) consumes the product or service. The level of (dis)satisfaction that a product or service user (the customer) would experience should not be expected of people who pay for a product or service but do not utilize it. In order to understand the concept of customer satisfaction, we must understand that it refers to user satisfaction rather than buyer happiness (which may include non-users) (Hom 2000). From now on, I will use “consumer satisfaction” instead of “customer satisfaction.” Feeling satisfied is something that, given a set of conditions, is a transient attitude that is easily changed (Hom 2000).

80 per cent of business problems can be solved through simple analysis methods such as: 1) Cumulative analysis 2) Correlation analysis 3) Trend analysis 4) estimation and generalization. In this study, we will help these organizations through customer satisfaction data, correlation analysis, and regression to discover precisely which areas they should take corrective measures in.

Regression Using data to discover the relationship between them is broadly called “data mining.” One of the tools for relationship measurement and modelling is the use of statistical regression tools. In order to analyze and discover the model of “data fog” (Big Data), different regression methods have been developed. The use of simple linear regression analysis is widely used in various data mining sciences, especially in the subject of “machine learning.” Linear regression is the first algorithm one would learn when beginning a career in machine learning or deep learning because it is simple to implement and apply in real-time. This algorithm is widely used in data science and statistical fields to model the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). Several types of regression techniques are available based on the data being used. Although linear regression involves simple mathematical logic, its applications are used across different fields in real-time (Kurama 2020).

Simple Linear Regression Simple linear regression is a statistical method that models the relationship between two variables, one independent (predictor) and one dependent (response). The goal is to fit a straight line to the data that best captures the relationship between the variables and predicts the response based on the predictor.

Multiple Linear Regression One of the conventional methods in multivariate analysis is the “multiple linear regression” technique. Regression analysis establishes a linear relationship between the “response variable” and one or more “explanatory variables.” Of course, sometimes, the response variable is called a “dependent variable,” and descriptive variables are also called “independent variables.” If the relationship between a response change and a change line is unchangeable, the regression technique is called simple linear regression. However, if several descriptive or independent variables are used in the regression model, the regression method is called “Multiple Linear Regression.” Of course, the non-differentiated regression method of multiple and independent response variables is also used, which is called “multivariate regression,” and more than one response variable is analyzed and modelled.

Regression Analysis Regression analysis estimates the relationship between a dependent variable and one or more independent variables. This technique is widely applied to predict the outputs by forecasting the data, analyzing the time series, and finding the causal effect dependencies between the variables. There are several regression techniques based on the number of independent variables, the dimensionality of the regression line, and the type of dependent variable. Out of these, the two most popular regression techniques are linear regression and logistic regression. Regression has numerous applications. Researchers use regression to indicate the strength of the impact of multiple independent variables on a dependent variable on different scales. For example, consider a data set of weather information recorded over the past few decades. We could use that data to forecast the weather for the next couple of years. Regression is also widely used in organizations and businesses to assess risk and growth based on previously recorded data (Kurama 2020).

The Math and Logic Behind Linear Regression The goal of linear regression is to identify the best-fit line passing through continuous data by employing a specific mathematical criterion. This technique falls under the umbrella of "supervised machine learning." Before jumping into linear regression, though, we should first understand what supervised learning is all about (Kurama 2020).

Applications of Linear Regression Linear regression is a powerful statistical technique that can generate insights into consumer behaviour, help to understand the business better, and comprehend factors influencing profitability. It can also be put to work evaluating trends and forecasting data in various fields. We can use linear regression to solve a few of our day-to-day problems related to supporting decision-making, minimizing errors, increasing operational efficiency, discovering new insights, and creating predictive analytics (Kurama 2020).

2 Data

solving business problems and can be used to answer random questions and explore relationships that take time to be intuitive. Information gathered from customer satisfaction surveys can also help your business stay current and better understand what customers want and need. The information was gathered from a survey of 305 people, in which respondents indicated how satisfied they were with the cars they had purchased. This survey asks 24 questions about different parts of the car, a general question about satisfaction and four other pieces of information, such as age, gender, level of education, and why the car was bought. The Iran Standard and Quality Inspection Company (ISQI) provided this information to me, which performs various tasks. One of the most important parts of the company is the customer and market research department, which also provides consulting services. Car manufacturers in Iran use these consultations and quality reports and sometimes by the Ministry of Industry, Mining, and Trade. This process of collecting information from customers is also done by questioners who are settled in the call centre of this company.

3 Survey

The best method for evaluating customer satisfaction is receiving direct customer feedback. The final questions depend on what you wish to know. As a result, we must create a survey that reflects the entire picture. Customer satisfaction surveys are essential for improving businesses and ensuring customers remain with them. A good customer satisfaction survey should provide the statistical data required for analyzing your stated objectives. These surveys can also help your business increase productivity and profitability by evaluating customer expectations of your products and services and their level of trust and loyalty to your company. This survey is designed based on the Likert scale and represents a 5-point description. The Likert scale is used to measure the point of view, feelings, judgment, and, in general, issues that are not visible but affect people's behaviour. We used the Likert scale in this questionnaire because we must have various answers to reflect all points of view when we want to measure people's satisfaction. These answers include very low, low, medium, high, and very high. A brief explanation of the Likert scale is below:

Likert Scale A Likert scale is a rating scale used in surveys and other research to measure attitudes, opinions, or perceptions. It is named after the American psychologist Rensis Likert, who developed the technique in the 1930s. The scale typically consists of a series of statements, each of which the respondent is asked to rate on a scale of agreement, usually ranging from strongly disagree to agree strongly. The scale can have a five or seven-point scale; it can have more or fewer points per the requirement. Likert scales are used in various research fields, including psychology, sociology, education, and marketing. They are considered reliable and valid methods of measuring attitudes and opinions and are widely used in survey research. I chose this scale for my survey because it can show customers' perceptions; all the information and figures can be seen, and you can see how the answers are distributed among the customers.

4 Purpose of Study

It is necessary to go beyond meeting users' basic demands for long-term business growth. Setting and measuring customer satisfaction goals is the best way to delight your customers and keep them returning over the long term. The achievement reflects a consumer's subjective perception of an organization or a product based on how well it matches his or her expectations of that organization or product. There are different dimensions of customer satisfaction in various organizations. For instance: Products: quality, lifetime, design, applicability, and performance Delivery: timely delivery and delivery speed Employees and services: availability of employees or representatives, knowledge of employees or representatives, speed of solving problems and dealing with complaints, speed of responding to items, after-sale services, and professional behaviour of employees or representatives. Competitive pricing; product value for price parity Organization: ease of communication, ease of business, and transparency In this regard, I want to answer these two questions:

Which part of the car has a more significant effect on overall satisfaction?

Which one of the factors has a more significant impact on consumer satisfaction?

5 Framework

Enhancing the quality of the car is possible by analyzing the data. For this purpose, car manufacturers should not only improve the parts of the car that received a lower score but also identify the parts of the car that consumers pay more attention to, or, in other words, parts that have the most significant effect on increasing the level of consumer satisfaction. Car manufacturers must also maintain the quality of those parts at an acceptable level or even improve them. The following list explains the procedure in further detail:

Logical Organization of the R Markdown Script

1. Install and load libraries of R
2. Import data
3. Process data
4. Produce outputs (tables, plots, etc.)
5. Save outputs, if applicable (.csv, .png, etc.)

5.1 Install and Load Libraries of R

In the first step, I need to install the packages for my project. So, I used the function install.packages() to install packages like tidyverse, ggplot, and likert, and then I used the library() function to load them. Tidyverse provides a consistent, easy-to-learn, and integrated approach to data analysis and manipulation. The packages in the tidyverse include ggplot2 for data visualization, dplyr for data manipulation, tidyr for data cleaning, and readr for data import, among others. The tidyverse aims to provide a set of tools that work together seamlessly and make data analysis more efficient and enjoyable.

```r
library("tidyverse")
```

This output is generated when the tidyverse package is loaded into R. It shows the version of each package within the tidyverse and confirms that they have been loaded successfully. The package versions listed indicate that you are using tidyverse version 1.3.2 and each of its sub-packages (e.g. dplyr version 1.0.10, tidyr version 1.2.1, etc.). The warning messages indicate that some of the packages in the tidyverse were built under a different version of R and may cause conflicts with other R functions with the same name. The Conflicts section shows any potential conflicts between the tidyverse packages and other packages in your environment.

```r
library("ggplot2")
library("likert")
```

5.2 Data Importing

I use the R programming language to analyze my data. I use confidential data that is not available from public sources. Before we can manipulate and analyze data with R, we must import data. R supports various file formats, including .docx, .xls, .txt, and comma-separated files like.csv.I imported MZ data with the function read.csv() and named the dataframe survey1.

```
survey1 <- read.csv("datascience1.csv", header = TRUE, sep = ",")
```

5.3 Data Cleaning

Initially, 305 people responded to the survey1. In this step, I defined a value named “respondent” and added five more respondents to my data. Because I wanted to make sure that there was a variety of 1, 2, 3, 4, and 5 values in answer to each question, this means all of the answers to the questions for respondent 306 would be 1, all of the answers to the questions for respondent 307 would be 2, and so on.

```
Respondent <- c("306", "307", "308", "309", "310")
Q01 <- c(1, 2, 3, 4, 5)
Q02 <- c(1, 2, 3, 4, 5)
Q03 <- c(1, 2, 3, 4, 5)
Q04 <- c(1, 2, 3, 4, 5)
Q05 <- c(1, 2, 3, 4, 5)
Q06 <- c(1, 2, 3, 4, 5)
Q07 <- c(1, 2, 3, 4, 5)
Q08 <- c(1, 2, 3, 4, 5)
Q09 <- c(1, 2, 3, 4, 5)
Q10 <- c(1, 2, 3, 4, 5)
Q11 <- c(1, 2, 3, 4, 5)
Q12 <- c(1, 2, 3, 4, 5)
Q13 <- c(1, 2, 3, 4, 5)
Q14 <- c(1, 2, 3, 4, 5)
Q15 <- c(1, 2, 3, 4, 5)
Q16 <- c(1, 2, 3, 4, 5)
Q17 <- c(1, 2, 3, 4, 5)
Q18 <- c(1, 2, 3, 4, 5)
Q19 <- c(1, 2, 3, 4, 5)
Q20 <- c(1, 2, 3, 4, 5)
Q21 <- c(1, 2, 3, 4, 5)
```

```
Q22 <- c(1, 2, 3, 4, 5)
Q23 <- c(1, 2, 3, 4, 5)
Q24 <- c(1, 2, 3, 4, 5)
Finalsatisfaction <- c(1, 2, 3, 4, 5)
```

Now there is fake data, including five fake respondents for 24 questions, and the final satisfaction of every customer, defined through the below function.

```
fake <- data.frame(Respondent,
    Q01, Q02, Q03, Q04, Q05, Q06, Q07, Q08, Q09, Q10,
    Q11, Q12, Q13, Q14, Q15, Q16, Q17, Q18, Q19, Q20,
    Q21, Q22, Q23, Q24, Finalsatisfaction)
```

I combined two data frames, survey1 and fake, by row using the rbind() function and named it survey2.

```
survey2 <- rbind(survey1, fake)
survey2$Q01_f <- as.factor(survey2$Q01)
survey2$Q02_f <- as.factor(survey2$Q02)
survey2$Q03_f <- as.factor(survey2$Q03)
survey2$Q04_f <- as.factor(survey2$Q04)
survey2$Q05_f <- as.factor(survey2$Q05)
survey2$Q06_f <- as.factor(survey2$Q06)
survey2$Q07_f <- as.factor(survey2$Q07)
survey2$Q08_f <- as.factor(survey2$Q08)
survey2$Q09_f <- as.factor(survey2$Q09)
survey2$Q10_f <- as.factor(survey2$Q10)
survey2$Q11_f <- as.factor(survey2$Q11)
survey2$Q12_f <- as.factor(survey2$Q12)
survey2$Q13_f <- as.factor(survey2$Q13)
survey2$Q14_f <- as.factor(survey2$Q14)
survey2$Q15_f <- as.factor(survey2$Q15)
survey2$Q16_f <- as.factor(survey2$Q16)
survey2$Q17_f <- as.factor(survey2$Q17)
survey2$Q18_f <- as.factor(survey2$Q18)
survey2$Q19_f <- as.factor(survey2$Q19)
survey2$Q20_f <- as.factor(survey2$Q20)
survey2$Q21_f <- as.factor(survey2$Q21)
survey2$Q22_f <- as.factor(survey2$Q22)
survey2$Q23_f <- as.factor(survey2$Q23)
survey2$Q24_f <- as.factor(survey2$Q24)
```

In this level, I defined a value named “factor levels,” ranging from very low to very high. As previously stated, very low satisfaction equals 1, and very high satisfaction equals 5.

```
factor_levels <- c("very low", "low", "medium", "high", "very high")

levels(survey2$Q01_f) <- factor_levels
levels(survey2$Q02_f) <- factor_levels
levels(survey2$Q03_f) <- factor_levels
levels(survey2$Q04_f) <- factor_levels
levels(survey2$Q05_f) <- factor_levels
levels(survey2$Q06_f) <- factor_levels
levels(survey2$Q07_f) <- factor_levels
levels(survey2$Q08_f) <- factor_levels
levels(survey2$Q09_f) <- factor_levels
levels(survey2$Q10_f) <- factor_levels
levels(survey2$Q11_f) <- factor_levels
levels(survey2$Q12_f) <- factor_levels
levels(survey2$Q13_f) <- factor_levels
levels(survey2$Q14_f) <- factor_levels
levels(survey2$Q15_f) <- factor_levels
levels(survey2$Q16_f) <- factor_levels
levels(survey2$Q17_f) <- factor_levels
levels(survey2$Q18_f) <- factor_levels
levels(survey2$Q19_f) <- factor_levels
levels(survey2$Q20_f) <- factor_levels
levels(survey2$Q21_f) <- factor_levels
levels(survey2$Q22_f) <- factor_levels
levels(survey2$Q23_f) <- factor_levels
levels(survey2$Q24_f) <- factor_levels
```

I used the “nrow” function to find the dimensions of my data frame and return the number of rows.

```
nrow(survey2)
```

```
## [1] 310
```

The “subset” function is used to choose a subset of a data frame based on specific criteria. I only wanted respondents who were less than 306. because the data from 306 to 310 were fake, and I added them myself.

```
survey3 <- subset(survey2, Respondent < 306)
nrow(survey3)
```

```
## [1] 230
```

The colnames() function is used to retrieve or set the column names of a data frame. When used without arguments, it returns the current column names of the specified data frame. When used with a character vector as an argument, it sets the column names of the specified data frame to the given character vector. In this section, I used this function to get the names.

```
colnames(survey3)
```

```
##  [1] "Respondent"        "Q01"               "Q02"
##  [4] "Q03"               "Q04"               "Q05"
##  [7] "Q06"               "Q07"               "Q08"
## [10] "Q09"               "Q10"               "Q11"
## [13] "Q12"               "Q13"               "Q14"
## [16] "Q15"               "Q16"               "Q17"
## [19] "Q18"               "Q19"               "Q20"
## [22] "Q21"               "Q22"               "Q23"
## [25] "Q24"               "Finalsatisfaction" "Q01_f"
## [28] "Q02_f"             "Q03_f"             "Q04_f"
## [31] "Q05_f"             "Q06_f"             "Q07_f"
## [34] "Q08_f"             "Q09_f"             "Q10_f"
## [37] "Q11_f"             "Q12_f"             "Q13_f"
## [40] "Q14_f"             "Q15_f"             "Q16_f"
## [43] "Q17_f"             "Q18_f"             "Q19_f"
## [46] "Q20_f"             "Q21_f"             "Q22_f"
## [49] "Q23_f"             "Q24_f"
```

```
survey4 <- survey3[,27:50]
colnames(survey4)
```

```
##  [1] "Q01_f" "Q02_f" "Q03_f" "Q04_f" "Q05_f" "Q06_f" "Q07_f" "Q08_f" "Q09_f"
## [10] "Q10_f" "Q11_f" "Q12_f" "Q13_f" "Q14_f" "Q15_f" "Q16_f" "Q17_f" "Q18_f"
## [19] "Q19_f" "Q20_f" "Q21_f" "Q22_f" "Q23_f" "Q24_f"
```

In this section, I defined the names of columns 1 to 24 as a VarHeading value. Then the names() function is used to retrieve or set the names of an object. It can be used on various objects, including vectors, lists, and data frames. As you can see, I used the VarHeading name I defined earlier in Survey 4. Now the data is ready to work with.

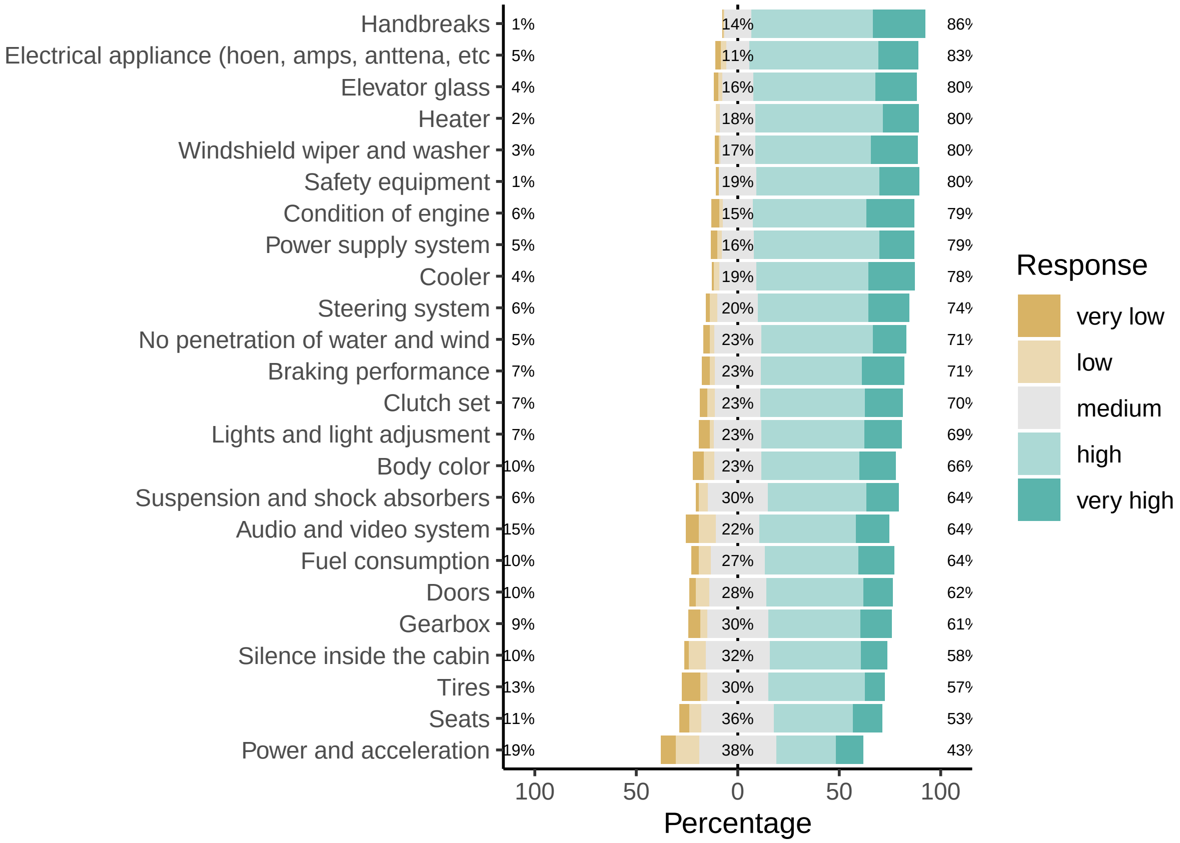
```
VarHeadings <- c("Doors",
"Body color","Silence inside the cabin",
"No penetration of water and wind","Power and acceleration","Fuel
consumption", "Condition of engine","Steering system","Clutch set","Gearbox",
"Suspension and shock absorbers", "Braking performance","Handbreaks",
"Windshield wiper and washer","Safety equipment","Seats","Tires","Lights and
light adjusment","Heater", "Cooler","Elevator glass","Power supply system",
"Electrical appliance (hoen, amps, anttena, etc","Audio and video system")

names(survey4) <- VarHeadings
colnames(survey4)
```

```
##  [1] "Doors"
##  [2] "Body color"
##  [3] "Silence inside the cabin"
##  [4] "No penetration of water and wind"
##  [5] "Power and acceleration"
##  [6] "Fuel\nconsumption"
##  [7] "Condition of engine"
##  [8] "Steering system"
##  [9] "Clutch set"
## [10] "Gearbox"
## [11] "Suspension and shock absorbers"
## [12] "Braking performance"
## [13] "Handbreaks"
## [14] "Windshield wiper and washer"
## [15] "Safety equipment"
## [16] "Seats"
## [17] "Tires"
## [18] "Lights and \nlight adjusment"
## [19] "Heater"
## [20] "Cooler"
## [21] "Elevator glass"
## [22] "Power supply system"
## [23] "Electrical appliance (hoen, amps, anttena, etc"
## [24] "Audio and video system"
```

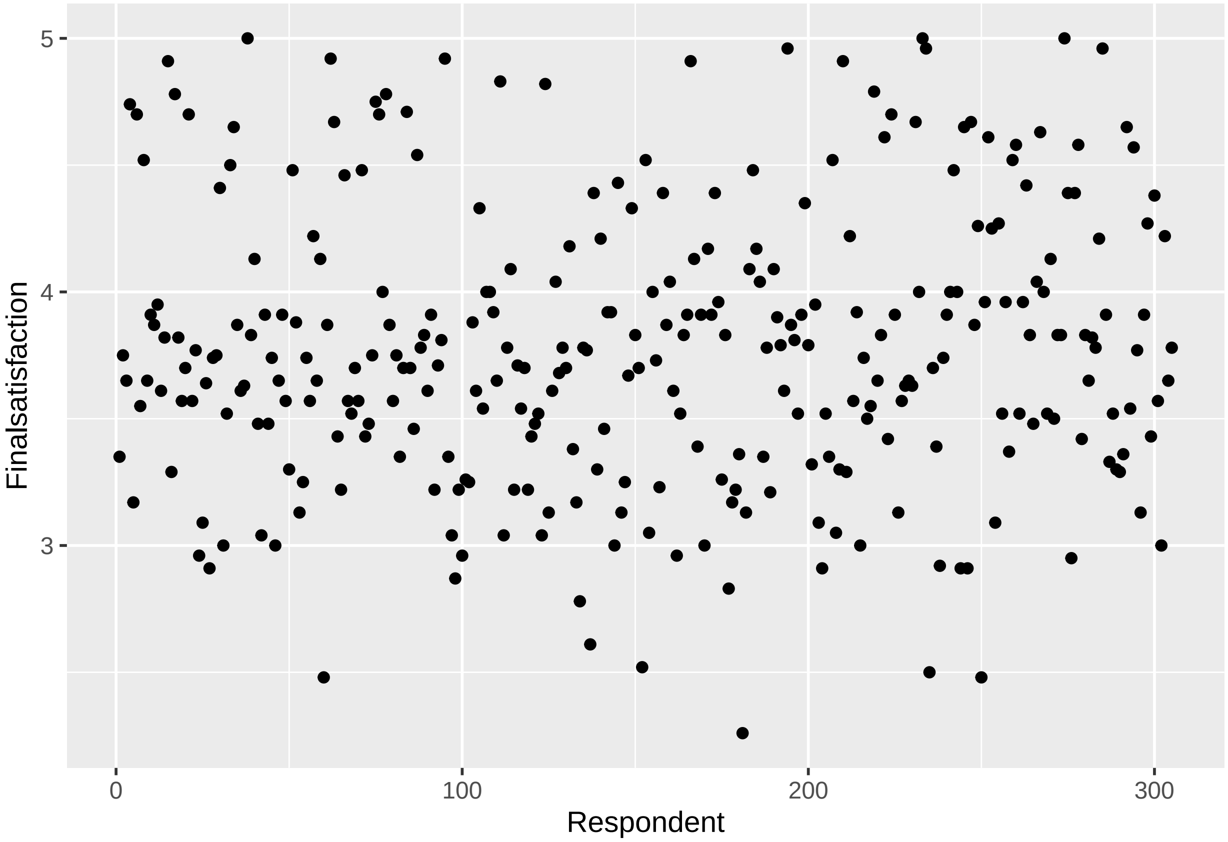
To display the following graph, I first called the Likert library and placed survey 4, which I had prepared in the previous steps. After adjusting the plot display settings and matching them, I finally used the plot function to display the graph. In this diagram, as you can see, the vertical vector includes twenty-four different parts of the car that customers are asked about. Moreover, this diagram shows that customer satisfaction in all parts tends to be satisfied or very satisfied. Furthermore, the level of satisfaction is categorized from high to low. For example, the level of satisfaction with the hand brake is higher than with the electrical appliance. As can be seen, 86% of the customers are satisfied with the performance of the handbrake, 14% have rated its performance as average or neutral, and only 1% of the interviewees were dissatisfied with its performance; these numbers for the electrical appliance are 83%, respectively. 83% are satisfied, 13% are neutral, and 5% are dissatisfied, a slightly higher percentage than HandBrake. In this way, we realize that we have severe problems in the power and acceleration section because less than 50% of the customers are satisfied with its performance.

```
library(likert)
p <- likert(survey4)
a <- likert.bar.plot(p, legend.position = "right", text.size = 2) +
   theme(text = element_text(size = rel(4)),axis.text.y = element_text(size=
   rel(2))) + theme_update(legend.text = element_text(size = rel(0.7))) +
   theme_classic()
plot(a)
```



At this stage, to show the dispersion of the data as much as possible and give an accurate view, we obtained a scatter plot by calling the ggplot2 library and the ggplot function. In this function, we set it to take the information from Survey 1 and set the X vector equal to the information in the Respondent column and the Y vector equal to the information in the Final Satisfaction column. As you can see, with as many as 305 interviewed customers, there are different points in this plot, each of which has a distinct final satisfaction, and this shows how scattered our statistical society is. Moreover, on the other hand, there is very little dispersion in the level of very high or very low scores, and most people gave average to good scores between 3 and 4.

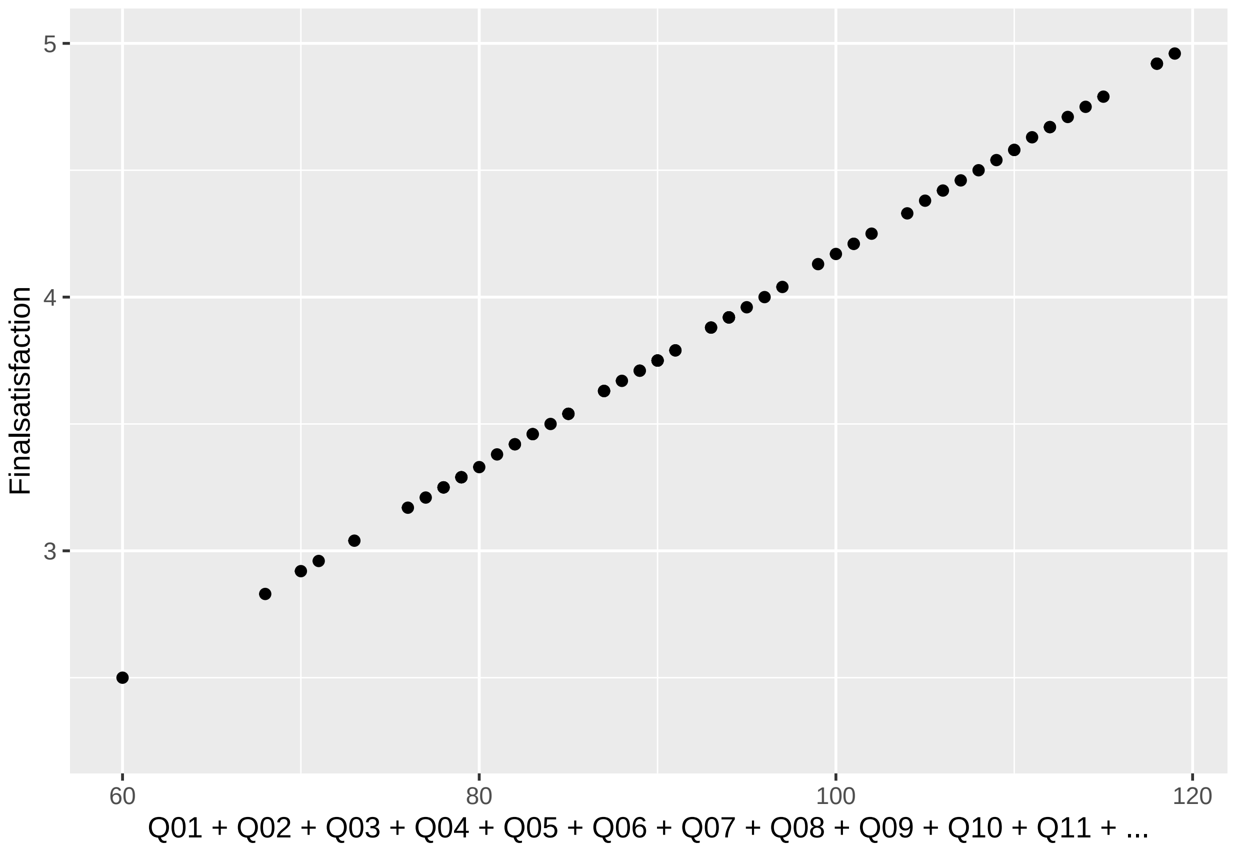
```
library(ggplot2)
ggplot(survey1, aes(x = Respondent, y = Finalsatisfaction)) + geom_point()
```



Using the ggplot library in R, this code makes a scatter plot with the sum of the variables Q01–Q24 on the x-axis and the final satisfaction on the y-axis. The data points are plotted as dots. From the shape of the line, it is understandable that there is a correlation between each factor and the final satisfaction because they are already a part of that. A warning message also indicates that 239 rows containing missing values have been removed from the data set before plotting a scatter plot using the geom_point() function. The warning informs the user that the data set used for the plot may not be complete and that some values have been removed. It is common for R to remove missing values before plotting to prevent errors or unexpected results.

```
library(ggplot2)
ggplot(survey1, aes(x = Q01+Q02+Q03+Q04+Q05+Q06+Q07+Q08+Q09+Q10+Q11+Q12+Q13
+Q14+Q15+Q16+Q17+Q18+Q19+Q20+Q21+Q22+Q23+Q24, y = Finalsatisfaction)) +
geom_point()
```

```
## Warning: Removed 239 rows containing missing values (`geom_point()`).
```

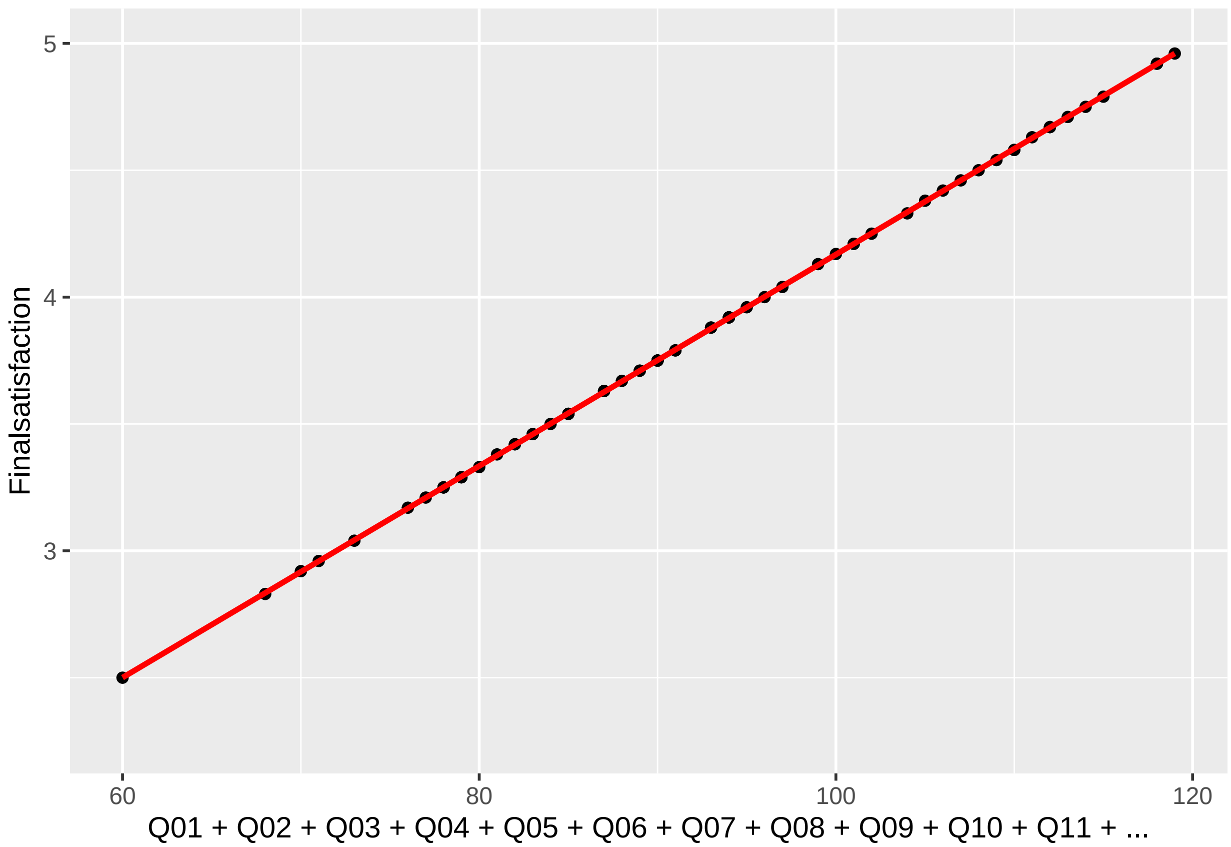


I put this code in place because I wanted to ensure that the line's slope was clear. I received warnings in this section then looked into it and discovered that these are warning messages generated by the R programming language when removing rows with non-finite or missing values while plotting a graph. The messages indicate that 239 rows with non-finite values or missing values were removed from the dataset used for plotting the graph. The stat_smooth() and geom_point() functions in R are commonly used for smoothing and plotting to scatter plots, respectively. However, in the end, I understood that these warnings would not cause a problem during the project, so I decided not to consider that or do anything special about it.

```
ggplot(survey1, aes(x=Q01+Q02+Q03+Q04+Q05+Q06+Q07+Q08+Q09+Q10+Q11+Q12+Q13
+Q14+Q15+Q16+Q17+Q18+Q19+Q20+Q21+Q22+Q23+Q24, y=Finalsatisfaction)) +
geom_point() + stat_smooth(formula=y~x, method="lm", se=FALSE, colour="red",
linetype=1)
```

```
## Warning: Removed 239 rows containing non-finite values (`stat_smooth()`).

## Warning: Removed 239 rows containing missing values (`geom_point()`).
```



I used the lm function, and this line fits a linear regression model with the response variable, Finalsatisfaction and predictor variables Q01 to Q24. The data for the regression is from the survey1 data frame. The ~ symbol indicates the relationship between the response and predictor variables. The fitted model is assigned to the variable model.

```
model  <- lm(Finalsatisfaction ~ Q01+Q02+Q03+Q04+Q05+Q06+Q07+Q08+Q09+Q10+Q11
+Q12+Q13+Q14+Q15+Q16+Q17+Q18+Q19+Q20+Q21+Q22+Q23+Q24, data = survey1 )
```

This line displays the summary of the linear regression model. This summary includes the coefficients, residuals, goodness of fit measures, etc. The show function in R displays a summary of the model object.

```
show(model)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q01 + Q02 + Q03 + Q04 + Q05 +
##     Q06 + Q07 + Q08 + Q09 + Q10 + Q11 + Q12 + Q13 + Q14 + Q15 +
##     Q16 + Q17 + Q18 + Q19 + Q20 + Q21 + Q22 + Q23 + Q24, data = survey1)
##
## Coefficients:
```

```
## (Intercept)         Q01         Q02         Q03         Q04         Q05
##   -0.004775    0.041226    0.041896    0.041297    0.041944    0.041283
##         Q06         Q07         Q08         Q09         Q10         Q11
##    0.041780    0.042028    0.041143    0.041325    0.042026    0.040833
##         Q12         Q13         Q14         Q15         Q16         Q17
##    0.041981    0.043267    0.041244    0.041382    0.042045    0.042743
##         Q18         Q19         Q20         Q21         Q22         Q23
##    0.041773    0.041195    0.042690    0.041690    0.040949    0.042745
##         Q24
##    0.040874
```

The first line retrieves the intercept term from the linear regression model and assigns it to the variable interm. The model$coefficients part gets the coefficients from the model object, and [1] picks the first element, which is the intercept term.

The second line retrieves the slope coefficient for the predictor variable Q01+Q02+Q03+Q04 +Q05+Q06+Q07+Q08+Q09+Q10+Q11+Q12+Q13+Q14+Q15+Q16+Q17+Q18+Q19 +Q20+Q21+Q22+Q23+Q24 and assigns it to the variable slope. The [2] part selects the second element, which is the slope coefficient.

The third line creates a new variable intercept by adding the intercept term to another coefficient in the linear regression model.

```
interm <- model$coefficients[1]
slope  <- model$coefficients[2]
interw <- model$coefficients[1]+model$coefficients[3]
```

A summary is a helpful tool for figuring out how to understand and explain the linear regression results. This line displays a summary of the linear regression model stored in the model object. The summary function provides a comprehensive summary of the model, including information about the coefficients, residuals, goodness of fit measures, etc.

```
summary(model)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q01 + Q02 + Q03 + Q04 + Q05 +
##     Q06 + Q07 + Q08 + Q09 + Q10 + Q11 + Q12 + Q13 + Q14 + Q15 +
##     Q16 + Q17 + Q18 + Q19 + Q20 + Q21 + Q22 + Q23 + Q24, data = survey1)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0050594 -0.0013457 -0.0001112  0.0012407  0.0051235
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0047751  0.0031545  -1.514    0.138
## Q01          0.0412257  0.0005404  76.291   <2e-16 ***
## Q02          0.0418965  0.0007050  59.426   <2e-16 ***
## Q03          0.0412969  0.0005988  68.966   <2e-16 ***
## Q04          0.0419444  0.0006412  65.412   <2e-16 ***
## Q05          0.0412828  0.0004518  91.367   <2e-16 ***
## Q06          0.0417797  0.0005690  73.423   <2e-16 ***
## Q07          0.0420277  0.0006893  60.971   <2e-16 ***
## Q08          0.0411428  0.0005215  78.899   <2e-16 ***
## Q09          0.0413252  0.0005819  71.018   <2e-16 ***
## Q10          0.0420263  0.0005473  76.782   <2e-16 ***
## Q11          0.0408331  0.0007568  53.956   <2e-16 ***
## Q12          0.0419811  0.0006073  69.127   <2e-16 ***
## Q13          0.0432674  0.0008829  49.007   <2e-16 ***
## Q14          0.0412444  0.0007659  53.849   <2e-16 ***
## Q15          0.0413821  0.0005269  78.537   <2e-16 ***
## Q16          0.0420447  0.0005403  77.824   <2e-16 ***
## Q17          0.0427431  0.0004204 101.667   <2e-16 ***
## Q18          0.0417727  0.0007486  55.798   <2e-16 ***
## Q19          0.0411950  0.0008894  46.320   <2e-16 ***
## Q20          0.0426896  0.0008567  49.832   <2e-16 ***
## Q21          0.0416898  0.0007157  58.249   <2e-16 ***
## Q22          0.0409486  0.0006409  63.896   <2e-16 ***
## Q23          0.0427447  0.0008656  49.379   <2e-16 ***
## Q24          0.0408741  0.0004706  86.847   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002587 on 41 degrees of freedom
##   (239 observations deleted due to missingness)
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.349e+05 on 24 and 41 DF,  p-value: < 2.2e-16
```

As I mentioned, this code creates a scatterplot using the ggplot function, with the data coming from the survey1 data frame. The x argument specifies the predictor variable (the sum of various survey question responses), and the y argument specifies the response variable (Finalsatisfaction).

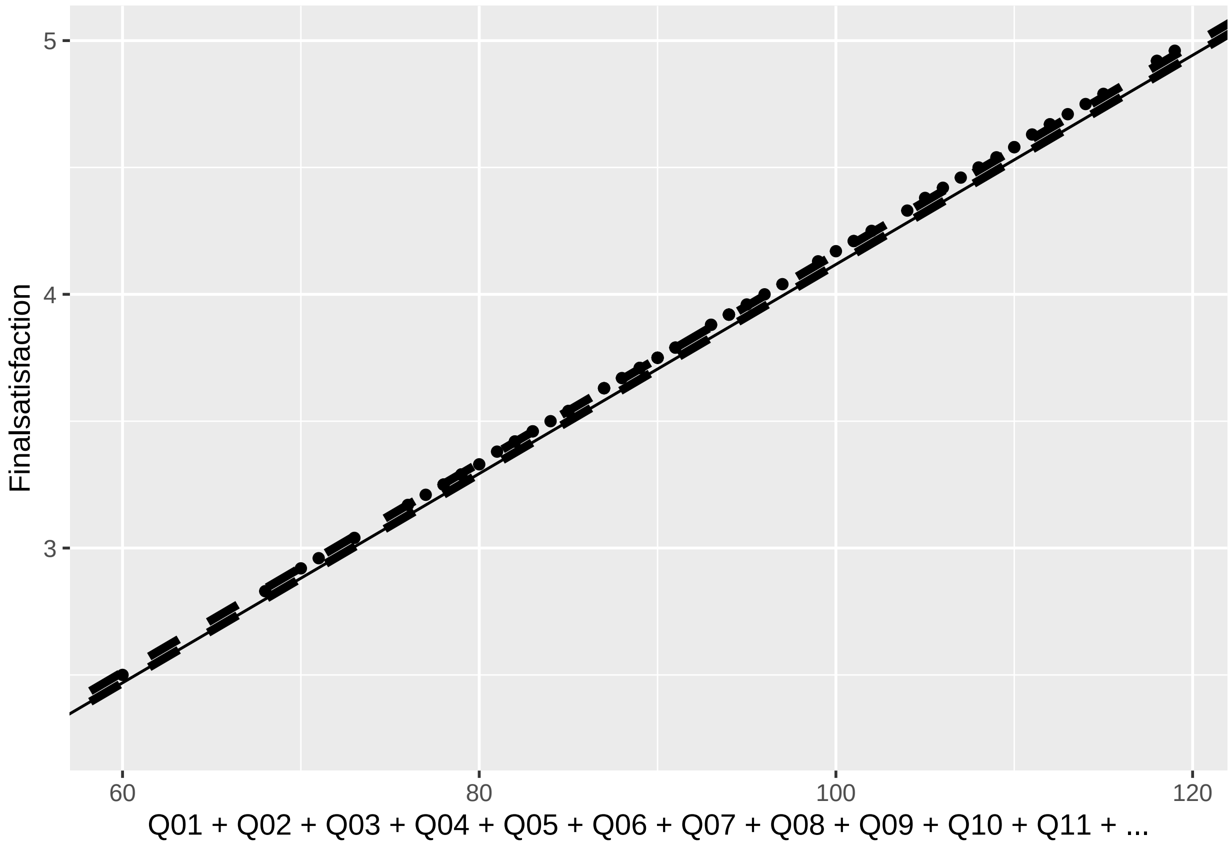
geom_point() adds a geom (geometric object) of type point to the plot, which displays individual data points as points on the scatterplot.

geom_abline (slope = slope, intercept = interw, linetype = 2, size = 1.5) adds a regression line to the plot, with a slope and intercept specified by the slope and interw variables, respectively. The linetype = 2 argument specifies that the line will be dotted, and size = 1.5 sets the size of the line.

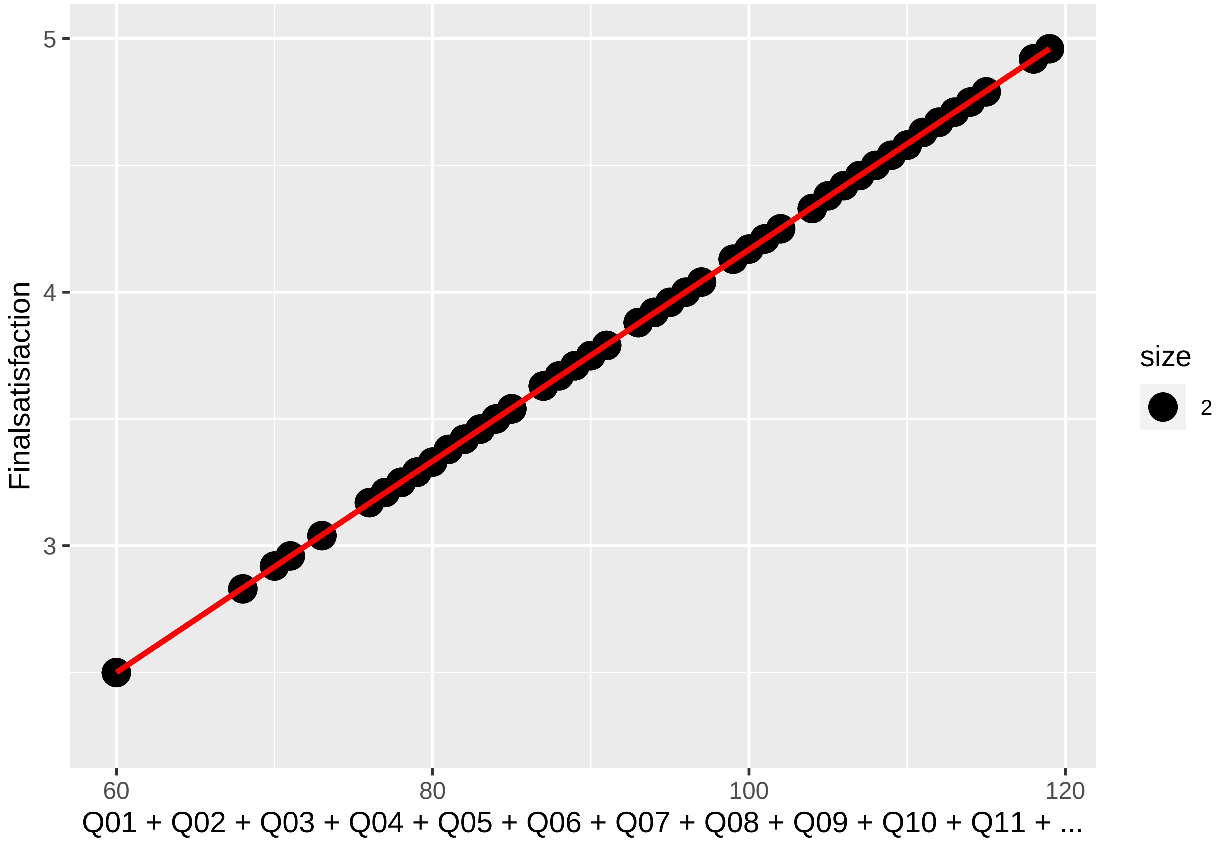
geom_abline(slope = slope, intercept = interm, linetype = 2, size = 1.5) is similar to the previous geom_abline, but uses the interm variable instead of interw to specify the intercept.

geom_abline: This code adds a third regression line to the plot (slope = coef(model)[[2]], intercept = coef(model)[[1]]), with the slope and intercept specified by the model object's coefficients. The coef function retrieves the coefficients from the model, and [[2]] and [[1]] select the slope and intercept, respectively.

```
ggplot(survey1, aes(x=Q01+Q02+Q03+Q04+Q05+Q06+Q07+Q08+Q09+Q10+Q11+Q12+Q13
+Q14+Q15+Q16+Q17+Q18+Q19+Q20+Q21+Q22+Q23+Q24, y=Finalsatisfaction)) +
geom_point() +
geom_abline(slope = slope, intercept = interw, linetype = 2, size=1.5)+
geom_abline(slope = slope, intercept = interm, linetype = 2, size=1.5) +
geom_abline(slope = coef(model)[[2]], intercept = coef(model)[[1]])
```

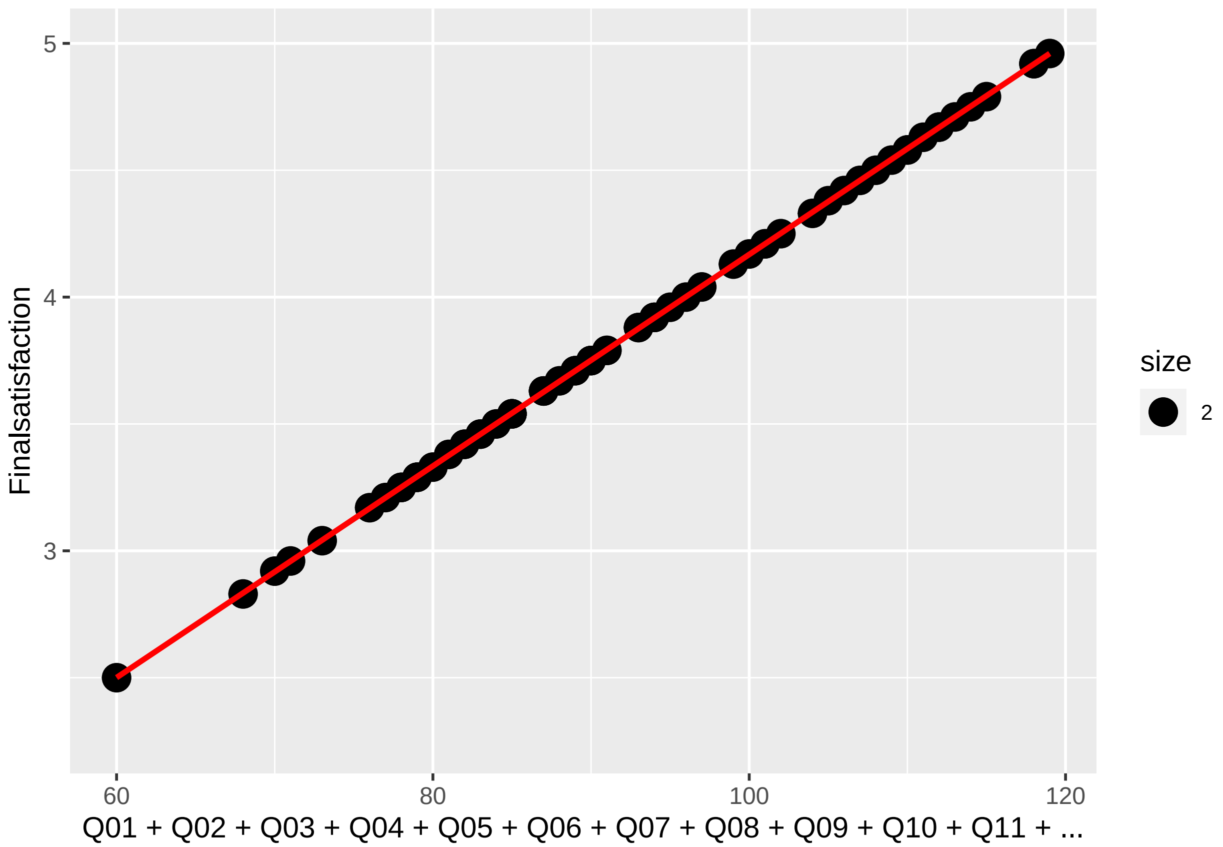


```
ggplot(survey1, aes(x=Q01+Q02+Q03+Q04+Q05+Q06+Q07+Q08+Q09+Q10+Q11+Q12+Q13
+Q14+Q15+Q16+Q17+Q18+Q19+Q20+Q21+Q22+Q23+Q24, y=Finalsatisfaction)) +
geom_point( aes(size = 2)) + stat_smooth(formula = y ~ x,  method = "lm",
se = FALSE, colour = "red", linetype = 1)
```



```
model  <- lm(Finalsatisfaction ~ Q01+Q02+Q03+Q04+Q05+Q06+Q07+Q08+Q09+Q10+Q11
+Q12+Q13+Q14+Q15+Q16+Q17+Q18+Q19+Q20+Q21+Q22+Q23+Q24 , data = survey1 )

ggplot(survey1, aes(x=Q01+Q02+Q03+Q04+Q05+Q06+Q07+Q08+Q09+Q10+Q11+Q12+Q13
+Q14+Q15+Q16+Q17+Q18+Q19+Q20+Q21+Q22+Q23+Q24, y=Finalsatisfaction)) +
geom_point( aes(size = 2)) +stat_smooth(formula = y ~ x, method = "lm",
se = T, colour = "red", linetype = 1)
```



This code performs a series of linear regression models (m1 to m5) where Finalsatisfaction is the dependent variable and Q01, Q02, Q01 * Q02, and Q03 are independent variables. The first model (m1) has only Q01 as the independent variable, the second (m2) has both Q01 and Q02, the third (m3) has both Q01 and Q02 and an interaction term between them, the fourth (m4) has Q01, Q02, an interaction term between them, and Q03. The fifth (m5) is similar to the third model but with the interaction term.

After fitting the models, the tab_model() function from the sjPlot library displays a table comparing the models, with p-values (indicating the significance of the independent variables) and coefficients presented. The p-values are presented with stars to indicate significance levels defined in p.threshold (0.2, 0.1, and 0.05). However, confidence intervals are not displayed (show.ci = FALSE), and standard errors are also not displayed (show.se = FALSE).

```r
m1 <- lm(Finalsatisfaction ~ Q01 , data = survey1 )
m2 <- lm(Finalsatisfaction ~ Q01 + Q02 , data = survey1 )
m3 <- lm(Finalsatisfaction ~ Q01 + Q02 + Q01 * Q02 , data = survey1 )
m4 <- lm(Finalsatisfaction ~ Q01 + Q02 + Q01 * Q02 + Q03 , data = survey1 )
m5 <- lm(Finalsatisfaction ~ Q01 + Q02 + Q01* Q02 , data = survey1)
```

This code creates a boxplot with jittered points to visualize the relationship between Q01 and Finalsatisfaction in the survey1 data.

The ggplot() function is used to create the plot, with Q01 as the x-axis, Finalsatisfaction as the y-axis, and Q01 defined as the grouping variable. The geom_boxplot() function is used to draw the boxplot and scale_fill_viridis() is used to set the color palette. The geom_jitter() function is used to add the jittered points with the specified color, size, and alpha level.

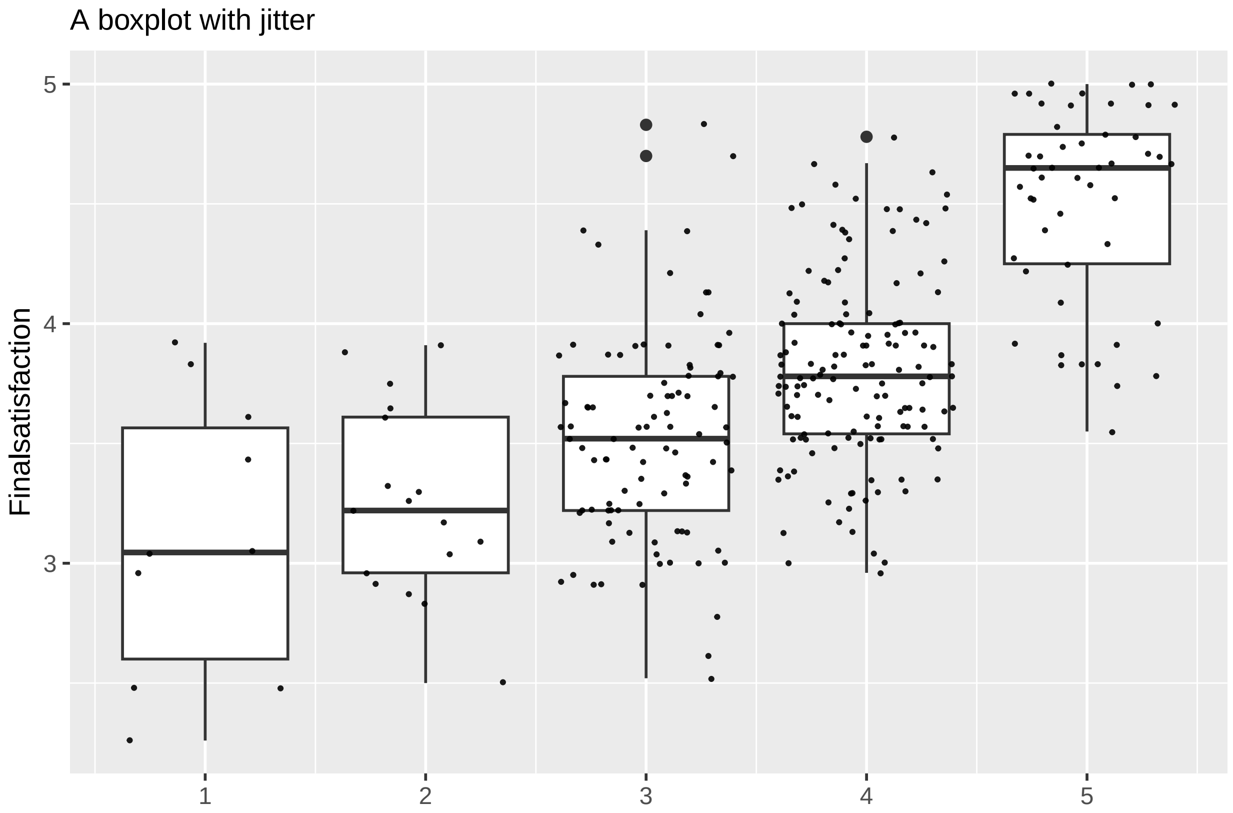
The theme() function is used to adjust the plot appearance, with the legend position set to "none" and the plot title size set to 11. The ggtitle() function is used to set the plot title, and xlab() is used to label the x-axis.

Loading the hrbrthemes and viridis libraries gives the plot more options for how to format it.

```
library(viridis)
```

```
## Loading required package: viridisLite
```

```
  ggplot(survey1, aes(x = Q01, y = Finalsatisfaction, group = Q01)) +
  geom_boxplot() +
  scale_fill_viridis(discrete = TRUE, alpha=0.6) +
  geom_jitter(color="black", size=0.4, alpha=0.9) +
  theme(
    legend.position="none",
    plot.title = element_text(size=11)
  ) +
  ggtitle("A boxplot with jitter") +
  xlab("")
```



6 Data Analysis

Using data to discover the relationship between them is the basis of data analysis. One of the tools for relationship measurement and modelling is regression. Regression is used to discover the model of a linear relationship between variables. In my case, one of the conventional methods is the "multiple linear regression" technique because I have 24 different variables. Regression analysis establishes a linear relationship between the "response variable" and one or more "explanatory variables." Of course, sometimes, the response variable is called a "dependent variable," and descriptive variables are also called "independent variables." I want to see which of the 24 possible answers to the project's 24 questions has the most significant impact on satisfaction. So, we need to make a 24-variable regression equation to answer this question. Which will be like this:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_{24} x_{i24} + \epsilon_i, i = 1, ..., n$$

When I tried to solve the equation, I found no way to get a result from a 24-variable regression equation due to the number of variables. Thus, I decided to solve the regression for each variable. In this respect, I had to run the code for each question and get the result, then compare the amount of R-squared. The stronger the correlation between that variable and final satisfaction, the higher the R-squared value.

```
model1 <- lm(Finalsatisfaction ~ Q01 , data = survey1)
model2 <- lm(Finalsatisfaction ~ Q02 , data = survey1)
model3 <- lm(Finalsatisfaction ~ Q03 , data = survey1)
model4 <- lm(Finalsatisfaction ~ Q04 , data = survey1)
model5 <- lm(Finalsatisfaction ~ Q05 , data = survey1)
model6 <- lm(Finalsatisfaction ~ Q06 , data = survey1)
model7 <- lm(Finalsatisfaction ~ Q07 , data = survey1)
model8 <- lm(Finalsatisfaction ~ Q08 , data = survey1)
model9 <- lm(Finalsatisfaction ~ Q09 , data = survey1)
model10 <- lm(Finalsatisfaction ~ Q10 , data = survey1)
model11 <- lm(Finalsatisfaction ~ Q11 , data = survey1)
model12 <- lm(Finalsatisfaction ~ Q12 , data = survey1)
model13 <- lm(Finalsatisfaction ~ Q13 , data = survey1)
model14 <- lm(Finalsatisfaction ~ Q14 , data = survey1)
model15 <- lm(Finalsatisfaction ~ Q15 , data = survey1)
model16 <- lm(Finalsatisfaction ~ Q16 , data = survey1)
model17 <- lm(Finalsatisfaction ~ Q17 , data = survey1)
model18 <- lm(Finalsatisfaction ~ Q18 , data = survey1)
model19 <- lm(Finalsatisfaction ~ Q19 , data = survey1)
model20 <- lm(Finalsatisfaction ~ Q20 , data = survey1)
model21 <- lm(Finalsatisfaction ~ Q21 , data = survey1)
model22 <- lm(Finalsatisfaction ~ Q22 , data = survey1)
model23 <- lm(Finalsatisfaction ~ Q23 , data = survey1)
model24 <- lm(Finalsatisfaction ~ Q24 , data = survey1)
summary(model1)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q01, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.01487 -0.31487 -0.02855  0.31145  1.29513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.41384    0.09976   24.20   <2e-16 ***
## Q01          0.37368    0.02652   14.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.429 on 303 degrees of freedom
## Multiple R-squared:  0.3959, Adjusted R-squared:  0.3939
## F-statistic: 198.6 on 1 and 303 DF,  p-value: < 2.2e-16
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q02, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.39403 -0.26683 -0.01963  0.25597  1.83915
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.45646    0.10190   24.11   <2e-16 ***
## Q02          0.35439    0.02648   13.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4381 on 301 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.3731, Adjusted R-squared:  0.3711
## F-statistic: 179.2 on 1 and 301 DF,  p-value: < 2.2e-16
```

```
summary(model3)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q03, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25749 -0.27851 -0.00851  0.28149  1.30251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.28442    0.09859   23.17   <2e-16 ***
## Q03          0.41102    0.02638   15.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4113 on 303 degrees of freedom
## Multiple R-squared:  0.4449, Adjusted R-squared:  0.4431
## F-statistic: 242.8 on 1 and 303 DF,  p-value: < 2.2e-16
```

```
summary(model4)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q04, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03113 -0.30552 -0.03113  0.26887  2.00571
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.31868    0.11434   20.28   <2e-16 ***
## Q04          0.38561    0.02939   13.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4382 on 291 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.3717, Adjusted R-squared:  0.3696
## F-statistic: 172.2 on 1 and 291 DF,  p-value: < 2.2e-16
```

```
summary(model5)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q05, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06631 -0.28631 -0.03377  0.25115  1.47861
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.72393    0.08155   33.40   <2e-16 ***
## Q05          0.31746    0.02342   13.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4355 on 303 degrees of freedom
## Multiple R-squared:  0.3774, Adjusted R-squared:  0.3754
## F-statistic: 183.7 on 1 and 303 DF,  p-value: < 2.2e-16
```

```
summary(model6)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q06, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.27606 -0.28106 -0.00132  0.27368  1.90445
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.50029    0.10532   23.74   <2e-16 ***
## Q06          0.34526    0.02764   12.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4482 on 293 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.3475, Adjusted R-squared:  0.3452
## F-statistic:   156 on 1 and 293 DF,  p-value: < 2.2e-16
```

```
summary(model7)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q07, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11345 -0.28757 -0.00757  0.31655  1.51417
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3841     0.1075   22.18   <2e-16 ***
## Q07           0.3559     0.0267   13.33   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4386 on 302 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3704, Adjusted R-squared:  0.3683
## F-statistic: 177.6 on 1 and 302 DF,  p-value: < 2.2e-16
```

```
summary(model8)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q08, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.55327 -0.29327  0.00718  0.30324  1.37021
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.19933    0.11338   19.40   <2e-16 ***
## Q08          0.40349    0.02833   14.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4272 on 303 degrees of freedom
## Multiple R-squared:  0.401,  Adjusted R-squared:  0.399
## F-statistic: 202.9 on 1 and 303 DF,  p-value: < 2.2e-16
```

```
summary(model9)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q09, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.20179 -0.28324  0.02176  0.27035  1.05676
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.25905    0.10020   22.55   <2e-16 ***
## Q09          0.39855    0.02557   15.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4114 on 302 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4459, Adjusted R-squared:  0.4441
## F-statistic:   243 on 1 and 302 DF,  p-value: < 2.2e-16
```

```
summary(model10)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q10, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99744 -0.27747 -0.02582  0.28087  1.55242
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.40096    0.09331   25.73   <2e-16 ***
## Q10          0.37662    0.02471   15.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4157 on 302 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4348, Adjusted R-squared:  0.433
## F-statistic: 232.4 on 1 and 302 DF,  p-value: < 2.2e-16
```

```
summary(model11)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q11, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.20073 -0.24552  0.00969  0.25948  1.05969
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.14200    0.10472   20.46   <2e-16 ***
## Q11          0.43958    0.02747   16.00   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.407 on 301 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.4597, Adjusted R-squared:  0.4579
## F-statistic: 256.1 on 1 and 301 DF,  p-value: < 2.2e-16
```

```
summary(model12)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q12, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.96934 -0.29895 -0.00895  0.29066  1.29145
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.27737    0.10869   20.95   <2e-16 ***
## Q12          0.39040    0.02758   14.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4283 on 303 degrees of freedom
## Multiple R-squared:  0.398,  Adjusted R-squared:  0.3961
## F-statistic: 200.4 on 1 and 303 DF,  p-value: < 2.2e-16
```

```
summary(model13)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q13, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.91966 -0.30354  0.00646  0.28257  1.49257
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.94908    0.16615   11.73   <2e-16 ***
## Q13          0.44611    0.04001   11.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.465 on 302 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2916, Adjusted R-squared:  0.2893
## F-statistic: 124.3 on 1 and 302 DF,  p-value: < 2.2e-16
```

```
summary(model14)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q14, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17510 -0.26278  0.02417  0.25722  1.19417
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.01352    0.12910   15.60   <2e-16 ***
## Q14          0.44232    0.03175   13.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4318 on 299 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.3936, Adjusted R-squared:  0.3916
## F-statistic: 194.1 on 1 and 299 DF,  p-value: < 2.2e-16
```

```
summary(model15)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q15, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30681 -0.26681 -0.00681  0.25319  2.13620
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.94279    0.14239   13.64   <2e-16 ***
## Q15          0.46100    0.03523   13.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4413 on 303 degrees of freedom
## Multiple R-squared:  0.361,  Adjusted R-squared:  0.3589
## F-statistic: 171.2 on 1 and 303 DF,  p-value: < 2.2e-16
```

```
summary(model16)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q16, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28351 -0.25378  0.00543  0.26569  1.12649
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.31035    0.08320   27.77   <2e-16 ***
## Q16          0.41106    0.02251   18.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3812 on 302 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5247, Adjusted R-squared:  0.5231
## F-statistic: 333.3 on 1 and 302 DF,  p-value: < 2.2e-16
```

```
summary(model17)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q17, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17614 -0.29462 -0.03462  0.30538  1.71081
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.79071    0.08936   31.23   <2e-16 ***
## Q17          0.28848    0.02487   11.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4565 on 295 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.3132, Adjusted R-squared:  0.3109
## F-statistic: 134.5 on 1 and 295 DF,  p-value: < 2.2e-16
```

```
summary(model18)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q18, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33822 -0.27917  0.01178  0.29272  1.22083
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.54199    0.10654   23.86   <2e-16 ***
## Q18          0.32906    0.02754   11.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4551 on 303 degrees of freedom
## Multiple R-squared:  0.3202, Adjusted R-squared:  0.318
## F-statistic: 142.7 on 1 and 303 DF,  p-value: < 2.2e-16
```

```
summary(model19)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q19, data = survey1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3297 -0.2487  0.0503  0.2114  1.0903
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.18552    0.31683   3.742 0.000381 ***
## Q19          0.66104    0.07791   8.484 3.23e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4195 on 67 degrees of freedom
##   (236 observations deleted due to missingness)
## Multiple R-squared:  0.5179, Adjusted R-squared:  0.5107
## F-statistic: 71.99 on 1 and 67 DF,  p-value: 3.225e-12
```

```
summary(model20)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q20, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28142 -0.28142  0.00858  0.26766  1.11858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11637    0.12388   17.09   <2e-16 ***
## Q20          0.42126    0.03072   13.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4325 on 300 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.3853, Adjusted R-squared:  0.3833
## F-statistic: 188.1 on 1 and 300 DF,  p-value: < 2.2e-16
```

```
summary(model21)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q21, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30926 -0.26926 -0.00926  0.30074  1.51809
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.19987    0.13850   15.88   <2e-16 ***
## Q21          0.39735    0.03428   11.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4594 on 303 degrees of freedom
## Multiple R-squared:  0.3072, Adjusted R-squared:  0.305
## F-statistic: 134.4 on 1 and 303 DF,  p-value: < 2.2e-16
```

```
summary(model22)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q22, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29606 -0.29606  0.00394  0.30919  1.52127
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.28140    0.12040   18.95   <2e-16 ***
## Q22          0.38366    0.03023   12.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4458 on 302 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3478, Adjusted R-squared:  0.3456
## F-statistic:   161 on 1 and 302 DF,  p-value: < 2.2e-16
```

```
summary(model23)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q23, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31534 -0.25534 -0.01534  0.25055  1.09393
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.96965    0.12100   16.28   <2e-16 ***
## Q23          0.45642    0.02998   15.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4158 on 302 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4343, Adjusted R-squared:  0.4324
## F-statistic: 231.8 on 1 and 302 DF,  p-value: < 2.2e-16
```

```
summary(model24)
```

```
##
## Call:
## lm(formula = Finalsatisfaction ~ Q24, data = survey1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41164 -0.27917 -0.02164  0.26706  1.44706
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.74671    0.09112   30.14   <2e-16 ***
## Q24          0.28623    0.02432   11.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4543 on 299 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.3165, Adjusted R-squared:  0.3143
## F-statistic: 138.5 on 1 and 299 DF,  p-value: < 2.2e-16
```

7 Conclusion

As you can see in the results that we got, the amount of R-squared is different, ranging from 0.3109 for Q17 to 0.5231 for Q16. This demonstrates that Q16, which refers to seat quality, is highly related to overall satisfaction. In other words, manufacturers can use this high correlation to improve final satisfaction with minor improvements in seat quality. In conclusion, this data science project aimed at analyzing the final satisfaction of car consumers. Our findings indicate a strong positive correlation between the two factors, with high-quality seats significantly predicting consumer Finalsatisfaction. These results provide valuable insights for car manufacturers, as they suggest that investing in seat quality can improve consumer Finalsatisfaction and ultimately increase brand loyalty. These findings also offer a valuable reference for consumers when making their next car purchase. Seat quality should be a significant consideration when evaluating a car's overall satisfaction.

Obstacles I Overcame

This project has some obstacles I can overcome but others that I cannot, as I will explain in the next chapter.

Data Quality The data I have collected may need to be completed, accurate, or consistent, making it difficult to conduct a reliable analysis. To address issues with data quality, I conducted a thorough data cleaning and preprocessing step. This may include checking for missing values, outliers, and inconsistencies and addressing them accordingly.

Lack of Domain Knowledge Because I was not familiar with the car industry, I needed help interpreting the data and identifying important features that influence customer satisfaction. However, I researched the car industry to overcome this obstacle and familiarized myself with critical concepts and terminology. I also got to consult with experts in the field to gain additional insights.

Overfitting With too many features or a complex model, I may fit the model to the noise in the data rather than the underlying pattern. In this regard, I limited the range of my data and decreased the number of variables to the most important of them.

Difficulty in Communicating Findings It may be challenging to present findings and recommendations clearly, and concisely which is understandable for people who are not experts in data science. This way, I used language and pictures that were easy to understand to explain my results and suggestions.

Remaining Challenges, Problems, and Weaknesses

Limited Data Size With only 305 data points, my sample size was too small to make accurate predictions or detect subtle patterns in the data. I could have thought about getting more data or using techniques like "data augmentation" to make my dataset bigger than it was, but I could not do that for several reasons, and this remained a weakness of my project.

Time Constraints I had limited time to complete the project, so I needed to prioritize and focus on the most critical tasks and impactful aspects of my project. This may have caused me to miss some information, but I set realistic goals and deadlines for each task to stay on track.

Difficulty in Choosing Fatures Because there are 24 different parts and variables in a car, it may be hard to figure out which features are the most important for predicting customer satisfaction.

Multicollinearity Refers to a situation where two or more predictor variables in a multiple regression model are highly correlated. This can cause problems in estimating the regression coefficients and interpreting the model results. When two or more predictor variables are highly correlated, they measure similar information and are redundant. This redundancy can lead to unstable and unreliable estimates of the regression coefficients, as small changes in the data can result in significant changes in the estimates. Additionally, when there is multicollinearity, it becomes difficult to determine the unique effect of each predictor variable on the response variable, as the effects of the predictor variables are confounded. There are a few common ways to detect multicollinearity in regression models. One way is to calculate the correlation matrix of the predictor variables and look for high correlation coefficients. Another way is to calculate the Variance Inflation Factor (VIF) for each predictor variable, which measures how much the coefficient estimate's variance is increased due to multicollinearity. VIF values greater than 1 indicate that there is multicollinearity present.

Reference List

Anderson, Eugene W., and Mary W. Sullivan. 1993. "The Antecedents and Consequences of Customer Satisfaction for Firms." *Marketing Science 12.2.*

Bloemer, José MM, and Jos GAM Lemmink. 1992. "The Importance of Customer Satisfaction in Explaining Brand and Dealer Loyalty." *Journal of Marketing Management.*

Fornell, Claes. 1992. "A National Customer Satisfaction Barometer: The Swedish Experience." *Journal of Marketing 56.1.*

Hom, Willard. 2000. "An Overview of Customer Satisfaction Models."

Kurama, Vihar. 2020. "A Guide to Logistic Regression with Tensorflow 2.0."

Oliver, Richard L. 1997. "CUSTOMER SATISFACTION RESEARCH." *The Handbook of Marketing Research: Uses, Misuses, and Future Advances, 1, 569-587.*

Wangenheim, Florian, and Tomas Bayon. 2004. "Satisfaction, Loyalty and Word of Mouth Within the Customer Base of a Utility Provider: Differences Between Stayers, Switchers and Referral Switchers." *Journal of Consumer Behaviour: An International Research Review 3.3.*