Advanced Data Analysis with R

ΕΡΓΑΣΙΕΣ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ

ANARGYROS TSADIMAS
AM: f3612318
Professor: Ioannis Ntzoufras

Table of Contents

Table of Figures and Tables

Abstract

This assignment is based on data that describe the rentals of SIXT, a member of the Motodynamics group. The data are provided by the Motodynamics group for use in the "Data Analysis" course of the M.Sc. in Statistics of AUEB. This assignment also constitutes the 1st Phase of the competition organized for AUEB students by Motodynamics in collaboration with the Department of Statistics

We are interested in describing the general customer profiles and understanding and predicting the on-desk total revenue of each rental in order to offer competitive prices and offers/discounts or increase the probability of a customer buying at a given cost.

1.Introduction

The file we analyze includes measurements for 5000 observations that each and every one of them describes a customer, using 50 variables that are explained in the table below(Table 1).

# of variables	*Variable's name*	*Type*	*Meaning*
1	***Res.no***	*character*	*reservation number*
2	***Agr..no***	*character*	*agreement number*
3	***Driver.ID***	*integer*	*driver's ID*
4	***Days***	*integer*	*reservation days*
5	***Agent.group***	*factor*	*booking source*
6	***Driver.Country_Disp***	*factor*	*country of origin of driver*
7	***Driver.Age***	*integer*	*driver's age*
8	***Pre.paid.Amount***	*numeric*	*prepaid amount*
9	***First Licence year***	*integer*	*first license year*
10	***Check out date***	*date*	*day the customer took the car*
11	***Check out Time***	*integer*	*time the customer took the car*
12	***Booking date***	*date*	*the customer booked the car*
13	***Booking time***	*integer*	*time the customer booked the car*
14	***AD***	*binary*	*additional driver*
15	***B***	*binary*	*gars fee*
16	***BC***	*binary*	*roadside protection*
17	***BE***	*binary*	*Loss damage waiver*
18	***BF***	*binary*	*Loss damage waiver minimum*
19	***BO***	*binary*	*booster seat*
20	***BR***	*binary*	*interior protection*
21	***BS***	*binary*	*baby seat*
22	***CS***	*binary*	*child seat*
23	***DI***	*binary*	*diesel engine*
24	***FDW***	*binary*	*full damage waiver*
25	***LD***	*binary*	*loss damage waiver*
26	***NV***	*binary*	*navigation system*
27	***PAI***	*binary*	*personal accident insurance*
28	***SC***	*binary*	*snow chains*
29	***SS***	*binary*	*seasonal supplement*

30	***SUB***	*binary*	*sub subscription*
31	***TF***	*binary*	*prepaid fuel*
32	***TG***	*binary*	*tyre and windscreen coverage*
33	***UPS***	*binary*	*upsell*
34	***Upgra***	*binary*	*upgrade car category*
35	***C/O Mileage***	*integer*	*took out mileage of the car*
36	***Check out Station***	*character*	*station ID*
37	***Group***	*factor*	*car group category*
38	***Charged group***	*factor*	*charged car group category*
39	***Internet Insurance Net Revenue***	*numeric*	*insurance products bought online*
40	***Internet Non Insurance Net Revenue***	*numeric*	*non insurance products bought online*
41	***Rental Cost Res***	*numeric*	*total rental coast reservation online*
42	***Sales Channel 2***	*factor*	*channel reservation came from*
43	***Segment***	*factor*	*type of reservation*
44	***Past Rentals Entry***	*integer*	*customer's bought history*
45	***Manufacturer***	*factor*	*car's manufacturer*
46	***Color***	*factor*	*car's color*
47	***Rate title***	*factor*	*reservation cost rate category*
48	***Status***	*factor*	*reservations status*
49	***OnDesk Insurance Net Revenue***	*numeric*	*insurance products bought on desk*
50	***OnDesk Non-Insurance Net Revenue***	*numeric*	*known insurance products bought on desk*

Table 1:Variables description

In this assignment we examine their relationship between the On Desk Revenue and all the other variables. We will perform descriptive analysis for the most important variables and pair wise associations between them. Finally ,we will construct a linear model that describes the data and also have predictive power.

2.Descriptive analysis and exploratory data analysis

In this section we will analyze and present our data. After we import them in R studio we will eliminate some of the observations that are considered missing or damaged values which leaves us with the final sample that consists of 3005 observations.

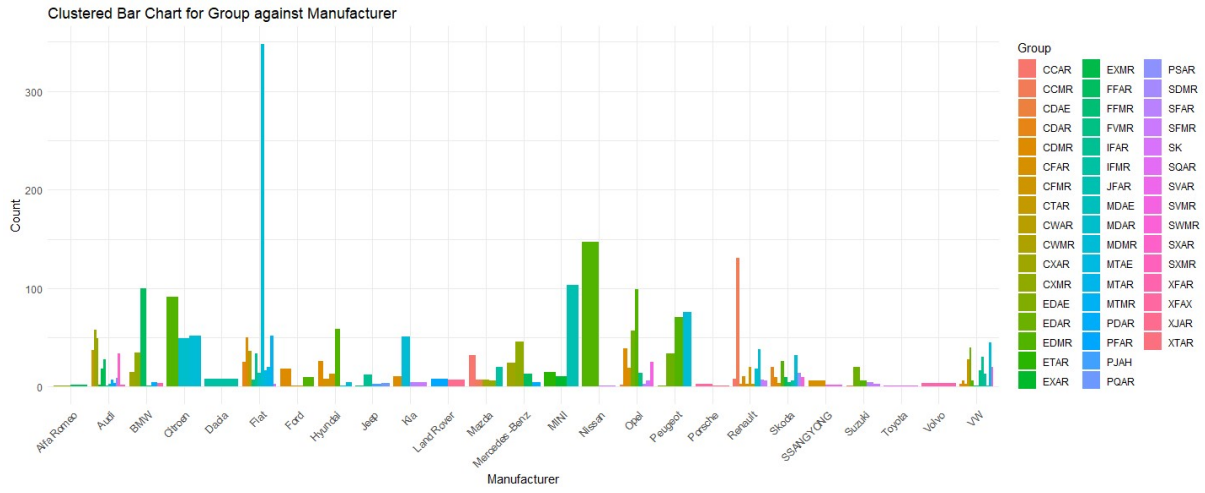

Figure 1:Barchart of group per Brand

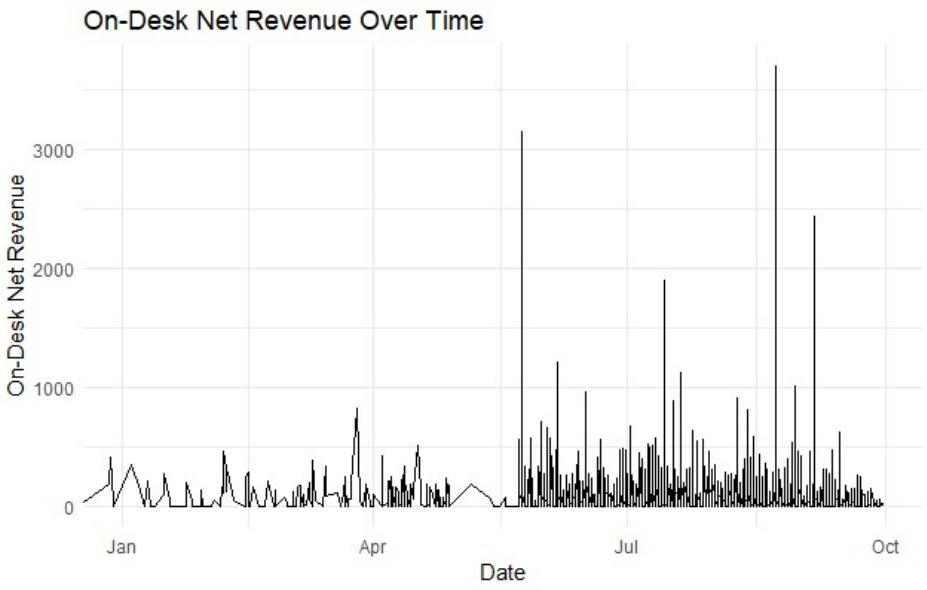

Figure 2:Net Revenue Across time

Here we're seeing not the most frequent rented cars are Fiat from the mini category and from the economy Nissan (Figure 1). We also see that the main bulk of our revenue came in the three summer months(Figure 2).

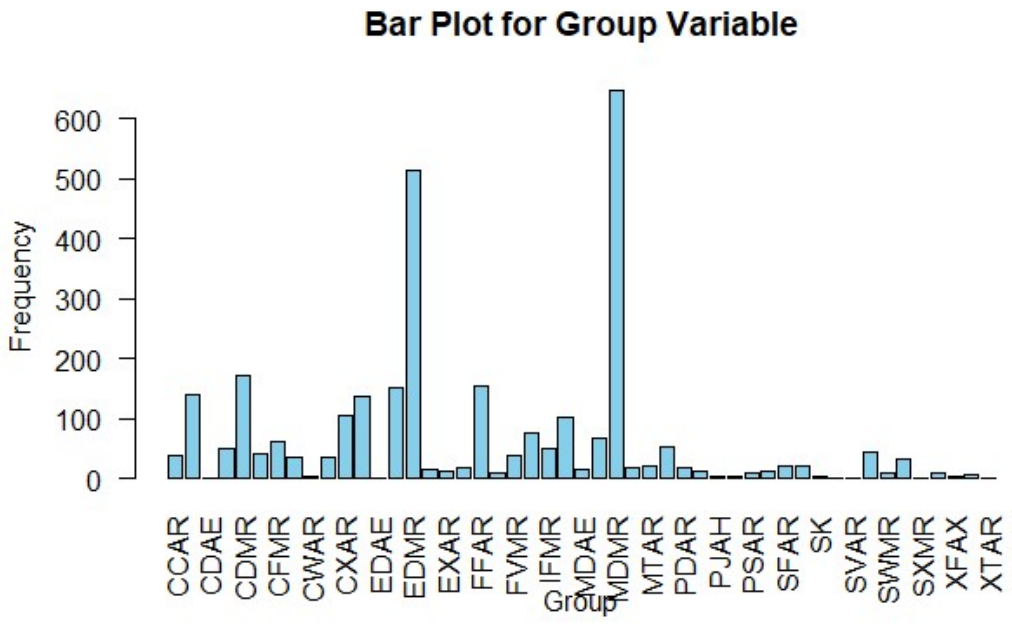

Figure 3:Popularity of groups

Most frequent rented categories are both mini and economy with 4/5 doors, manual and with air conditioning (Figure 3).

3.Pairwise comparisons

In this section of the report we will conduct pairwise comparisons between the variables to further analyze our data and draw better conclusions.

Below we have a correlation matrix in order to visualize the relationships between numeric variables (Figure 4).

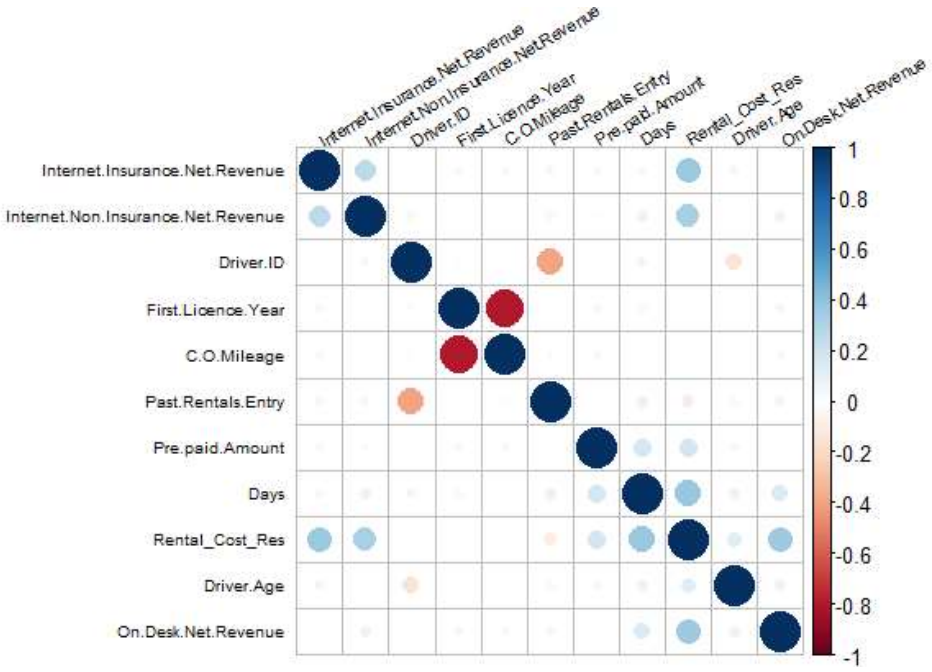

Figure 4:Correlation matrix for numeric variables

Here we see that only Volvo has higher revenue mean across all manufacturers.

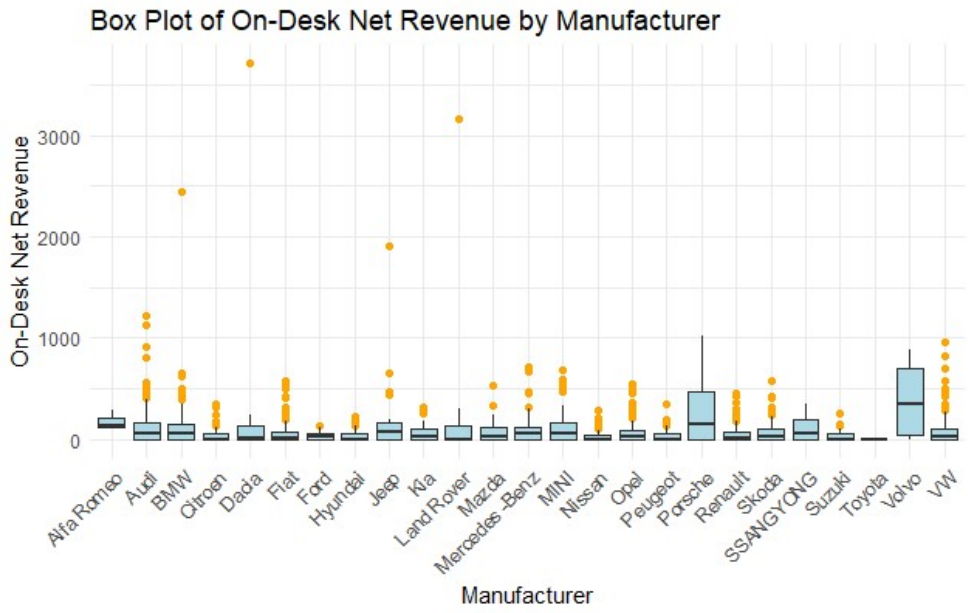

Figure 5:Revenue per brand

4.Predictive and Descriptive models

Now that we have analyzed our variables we can proceed to construct our model.

We start from the full model, the one that takes into account all the variables and examine which of them appear significant (R-squared= 0.49). In the ANOVA analysis only 17 of them appear to be statistically significant.Then using LASSO method for minimum lamba (3.26) 18 variables and the we used stepwise method to leaves as with just 13 variables.After that I excluded variables that appeared to be statistically insignificant which left us with the best model m6.

$$On.Desk.Net.Revenue=\beta0(0.83)+(LD)+\beta1(Group)+\beta2(Charged.group)+\beta3(Rental\_Cost\_Res)+\beta4(Manufacturer)+\epsilon\ ,\ \varepsilon\sim(0,\ 113.4)$$

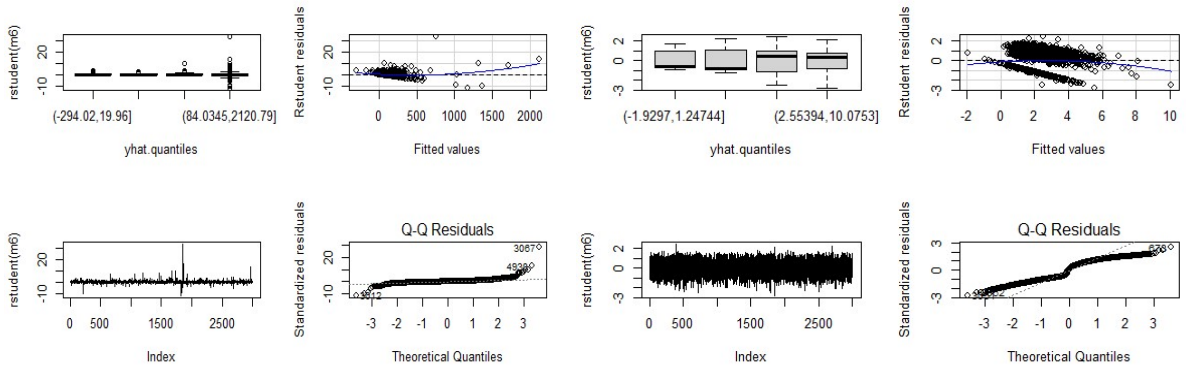

Figure 6:assumptions for the model without and with log transformation in response

Then we tested this on our testing dataset, with similar results.

5.Conclusions

After concluding the analysis, we can make several observations. Factors such as the customer's group car category choice, total reservation online cost and the manufaturer play important roles in determining whether a customer will buy on desk products. This information can be useful for predicting or influencing driver's behavious with targeted offers . Further research and analysis in the future could strengthen and validate our findings.

6.References

[1] Ntzoufras I. ,(2023) Advanced data analysis with R, educational notes for MSc program Statistics AUEB