REPORT SERIES WITH DLOOKR

Exploratory Data Analysis Report

Author:
dlookr package

Version:
0.4.0

March 11, 2021

Contents

Chapter 1

Introduction

The EDA Report provides exploratory data analysis information on objects that inherit data.frame and data.frame.

1.1 Information of Dataset

The dataset that generated the EDA Report is an 'data.frame' object. It consists of 359,392 observations and 9 variables.

1.2 Information of Variables

Table 1.1: Information of Variables

variables	types	missing_count	missing_percent	unique_count	unique_rate
Payment_Mode	character	0	0	2	0.000
Company	character	0	0	2	0.000
City	character	0	0	19	0.000
KM Travelled	numeric	0	0	874	0.002
Price Charged	numeric	0	0	99176	0.276
Gender	factor	0	0	2	0.000
Income (USD/Month)	numeric	0	0	22725	0.063
Profit_margin	numeric	0	0	301825	0.840
Age_Group	factor	0	0	7	0.000

The target variable of the data is 'Profit_margin', and the data type of the variable is numeric.

1.3 About EDA Report

EDA reports provide information and visualization results that support the EDA process. In particular, it provides a variety of information to understand the relationship between the target variable and the rest of the variables of interest.

Chapter 2

Univariate Analysis

2.1 Descriptive Statistics

edaData
9 Variables 359392 Observations

Payment_Mode

n	missing	distinct
359392	0	2

```
Value        Card   Cash
Frequency  215504 143888
Proportion    0.6    0.4
```

Company

n	missing	distinct
359392	0	2

```
Value        Pink Cab Yellow Cab
Frequency       84711     274681
Proportion      0.236      0.764
```

City

n	missing	distinct
359392	0	19

```
lowest : ATLANTA GA     AUSTIN TX      BOSTON MA      CHICAGO IL     DALLAS TX
highest: SAN DIEGO CA   SEATTLE WA     SILICON VALLEY TUCSON AZ      WASHINGTON DC
```

KM Travelled

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
359392	0	874	1	22.57	14.12	3.57	5.80	12.00	22.44	32.96	39.20	42.00

```
lowest :  1.90  1.92  1.94  1.96  1.98, highest: 46.41 46.80 47.20 47.60 48.00
```

Price Charged

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
359392	0	99176	1	423.4	303.4	63.42	99.23	206.44	386.36	583.66	792.79	944.89

```
lowest :   15.60   15.75   16.38   16.53   16.76, highest: 1981.05 1993.83 2013.95 2016.70 2048.03
```

Gender

n	missing	distinct
359392	0	2

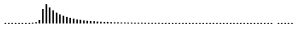
```
Value      Female   Male
Frequency  153480 205912
Proportion  0.427  0.573
```

Income (USD/Month)

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
359392	0	22725	1	15049	9104	3245	4525	8424	14685	21035	24793	29659

```
lowest :  2000  2001  2002  2003  2004, highest: 34985 34989 34995 34996 35000
```

Profit_margin



```
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
  359392        0   301825        1    137.3      159   -5.083    5.207   28.012   81.962
     .75      .90      .95
 190.030  358.985  478.564

lowest : -220.0600 -198.6980 -176.9308 -168.9850 -164.0400
highest: 1408.3440 1424.1408 1433.3420 1445.2720 1463.9660
```

Age_Group

n	missing	distinct
359392	0	7

```
lowest : 18-25 26-32 33-39 40-46 47-53, highest: 33-39 40-46 47-53 54-60 61+

Value      18-25 26-32 33-39 40-46 47-53 54-60   61+
Frequency  93344 79577 78681 35072 27390 26417 18911
Proportion 0.260 0.221 0.219 0.098 0.076 0.074 0.053
```

2.2 Normality Test of Numerical Variables

2.2.1 Statistics and Visualization of (Sample) Data

KM Travelled

* normality test : Shapiro-Wilk normality test
- statistic : 0.96501, p-value : 3.52992E-33

Table 2.1: skewness and kurtosis : KM Travelled

type	skewness	kurtosis
original	0.0707	1.9059
log transformation	-1.0642	3.4564
sqrt transformation	-0.4135	2.1874

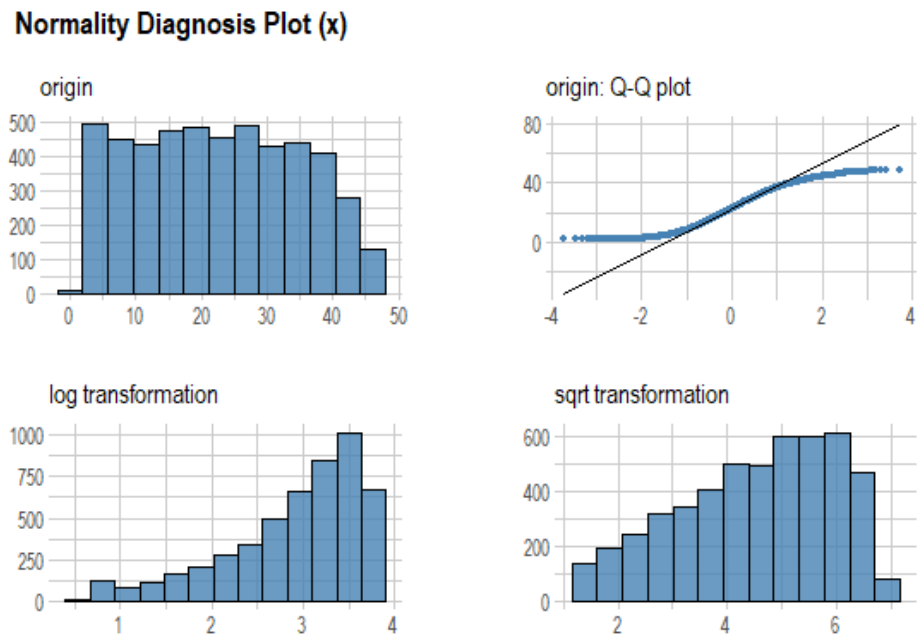


Figure 2.1: KM Travelled

Price Charged

* normality test : Shapiro-Wilk normality test
- statistic : 0.94705, p-value : 4.10616E-39

Table 2.2: skewness and kurtosis : Price Charged

type	skewness	kurtosis
original	0.8737	3.7984
log transformation	-0.8414	3.3380
sqrt transformation	0.0567	2.5414

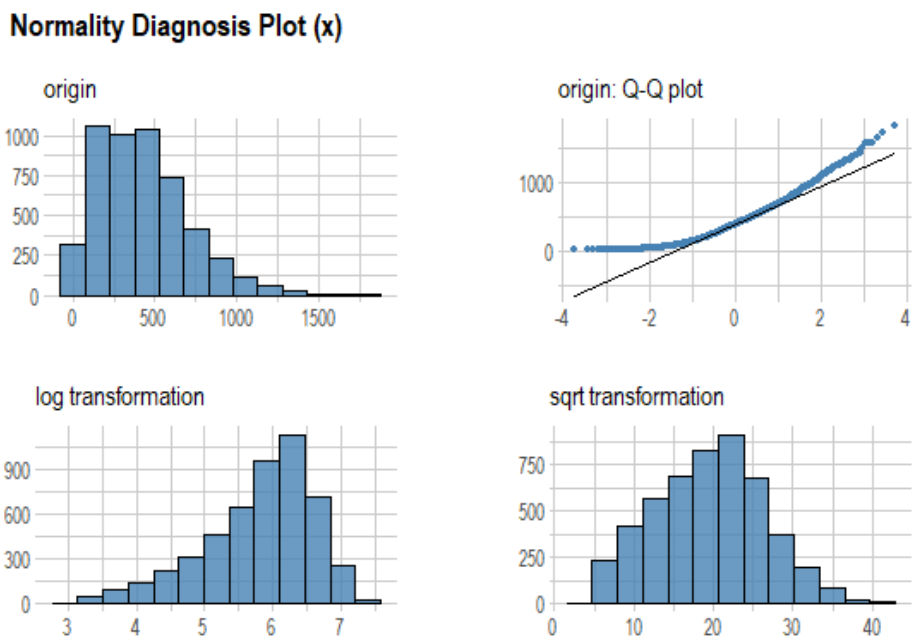


Figure 2.2: Price Charged

Income (USD/Month)

* normality test : Shapiro-Wilk normality test
- statistic : 0.97273, p-value : 7.37304E-30

Table 2.3: skewness and kurtosis : Income (USD/Month)

type	skewness	kurtosis
original	0.3246	2.3791
log transformation	-0.8011	2.9313
sqrt transformation	-0.2055	2.2337

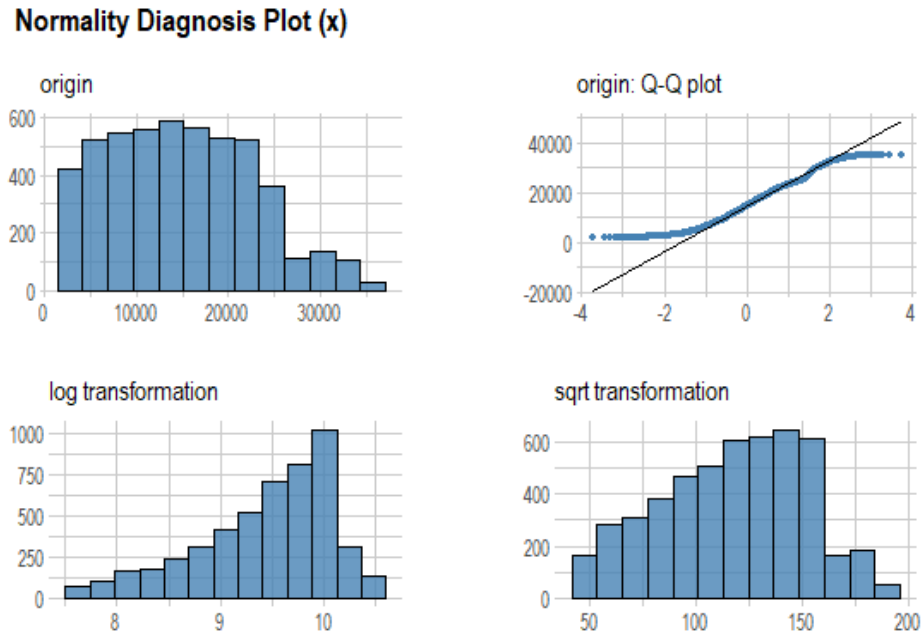


Figure 2.3: Income (USD/Month)

Chapter 3

Relationship Between Variables

3.1 Correlation Coefficient

3.1.1 Correlation Coefficient by Variable Combination

Table 3.1: The correlation coefficients (0.5 or more)

Variable1	Variable2	Correlation Coefficient
Profit_margin	Price Charged	0.864
Price Charged	KM Travelled	0.836

3.1.2 Correlation Plot of Numerical Variables

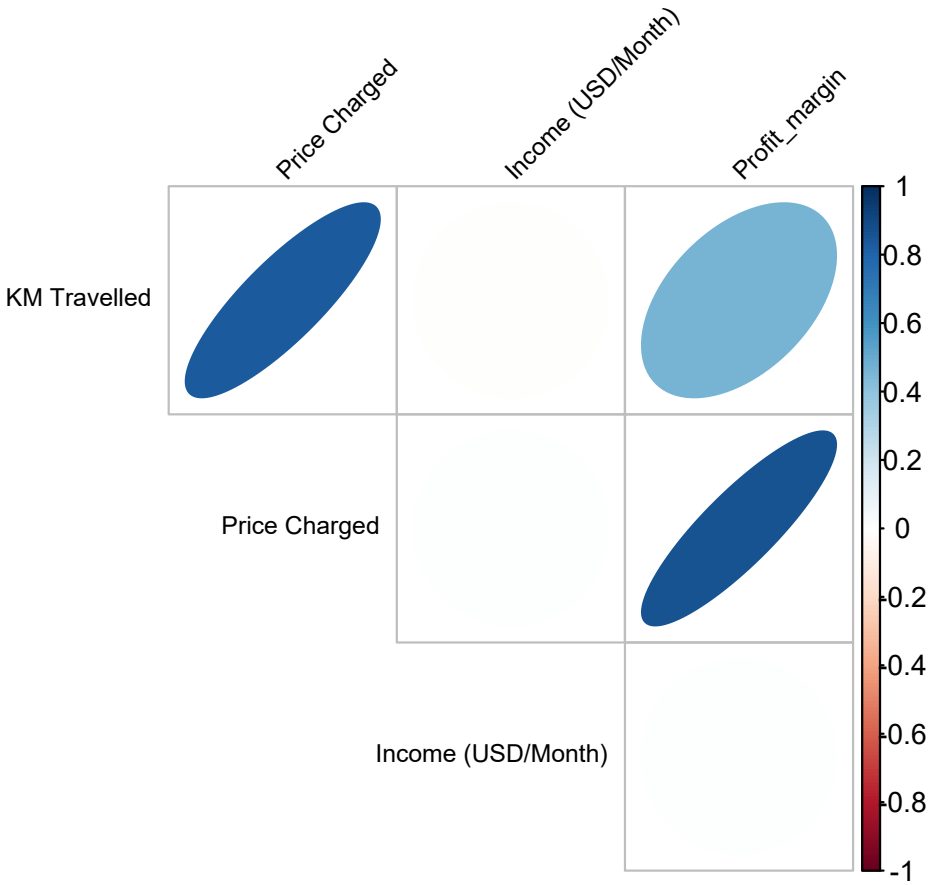


Figure 3.1: The correlation coefficient of numerical variables

Chapter 4

Target based Analysis

4.1 Grouped Descriptive Statistics

4.1.1 Grouped Numerical Variables

KM Travelled

```
## Error in str2lang(x):  <text>:1:20:  unexpected symbol
## 1:  Profit_margin ~ KM Travelled
##                        ^
```

4.1.2 Grouped Categorical Variables

Gender

1. **Analysis of Variance**

Table 4.1: Analysis of Variance Table : Gender

	Df	Sum Sq	Mean Sq	F value	Pr(> | F |)
Gender	1	4144145	4144144.97	161.32	0
Residuals	359390	9232163555	25688.43	NA	NA

2. **Simple Linear Model Information**

Residual standard error: 160 on 359390 degrees of freedom
Multiple R-squared: 0.00045, Adjusted R-squared: 0.00045
F-statistic: 161 on 1 and 359390 DF, p-value: 0

Table 4.2: Simple Linear Model coefficients : Gender

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	133.32	0.41	325.88	0
GenderMale	6.86	0.54	12.70	0

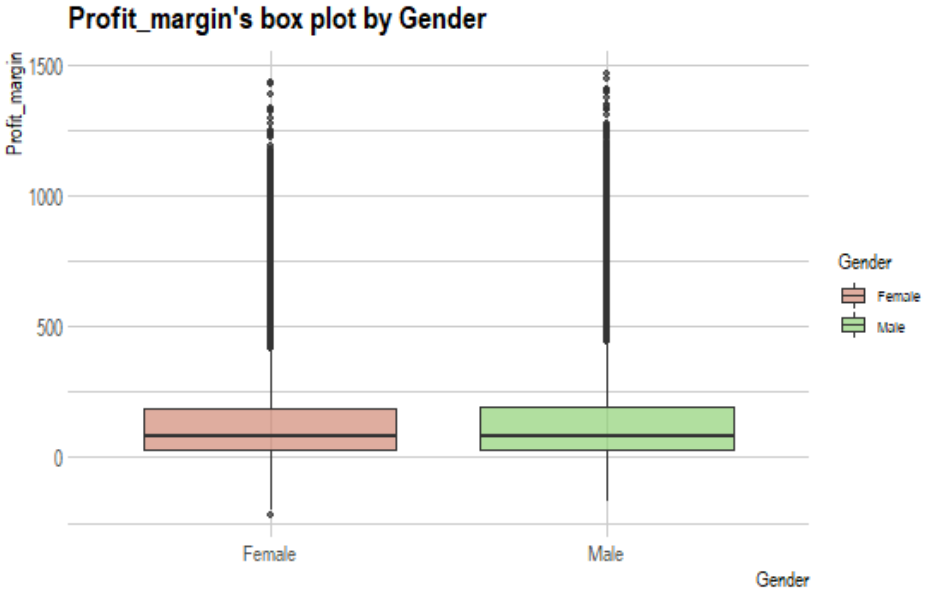


Figure 4.1: Gender

Age_Group

1. Analysis of Variance

Table 4.3: Analysis of Variance Table : Age_Group

	Df	Sum Sq	Mean Sq	F value	Pr(> | F |)
Age_Group	6	1839126	306521.0	11.93	0
Residuals	359385	9234468574	25695.2	NA	NA

2. Simple Linear Model Information

Residual standard error: 160 on 359385 degrees of freedom
Multiple R-squared: 2e-04, Adjusted R-squared: 0.00018
F-statistic: 12 on 6 and 359385 DF, p-value: 0.1734602

Table 4.4: Simple Linear Model coefficients : Age_Group

	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	137.66	0.52	262.37	0.00
Age_Group26-32	-1.05	0.77	-1.36	0.17
Age_Group33-39	0.47	0.78	0.61	0.54
Age_Group40-46	-0.52	1.00	-0.52	0.61
Age_Group47-53	4.82	1.10	4.37	0.00
Age_Group54-60	-3.12	1.12	-2.80	0.01
Age_Group61+	-6.86	1.28	-5.36	0.00

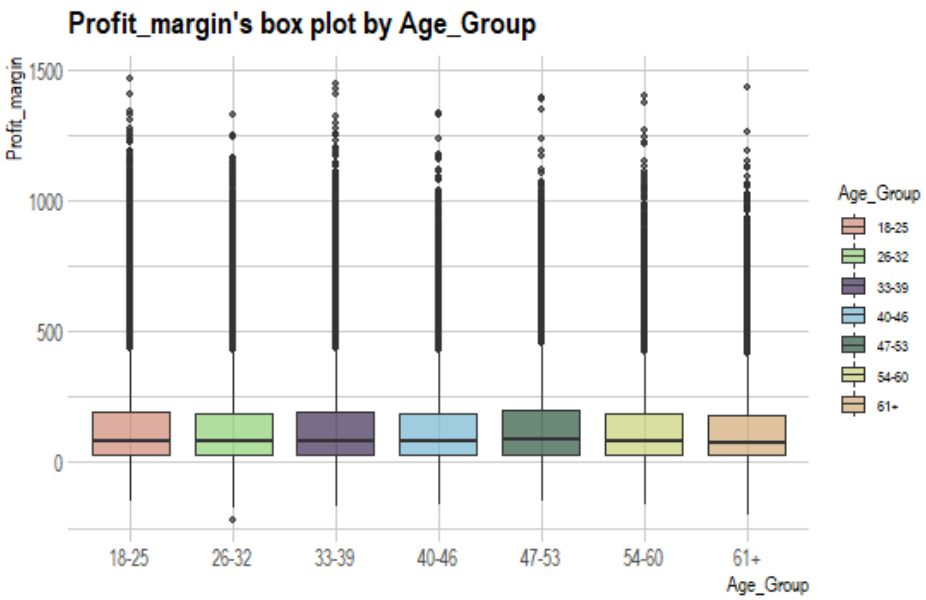


Figure 4.2: Age_Group

4.2 Grouped Relationship Between Variables

4.2.1 Grouped Correlation Coefficient

Numerical target variables are not supported.

4.2.2 Grouped Correlation Plot of Numerical Variables

Numerical target variables are not supported.