Table of Contents

Introduction

The report aims to conduct descriptive analysis, develop and interpret a multivariate linear regression model using the PortTravel dataset.

By the end of this report we would have accomplished the following:

1. Descriptive Analysis of at least 10 variables.
2. A comprehensive correlation table and highlighting how the correlation of the no_of_trips variable with other variables.
3. Develop a multivariate linear regression model, discussing the development process and techniques.
4. Model Description and interpretation.
5. Performing predictions using our model.

Assessing the data set

The travel dataset comprises 16 columns and 4,888 rows. Initial examination reveals 12 numeric columns and 4 character columns. It's essential to verify whether certain numeric columns represent categories. Table 2.1 shows the correct variable types and their representation in the plots.

Figure 1 highlights the columns with missing values, deferred for handling during the later stages of model development to minimise their influence on prior analysis. Figure 2 attempts to provide a visual depiction of the location of these missing values in the dataset. The columns are renamed for better reference and a neater plot label. Their arrangement follows a sequence from the highest to lowest missingness, aiding in detecting potential overlaps across columns, though few overlaps are observed in the plot.

Figure 3 highlights two columns with potential outliers. Buxton et al. (2003) describe the IQR method as a prevalent approach for outlier detection, outlined by the formula provided below.

Inter-Quartile Range(IQR) = Q3 – Q1

Lower Threshold = Q1 – (1.5xIQR)

Upper Threshold = Q3 + (1.5xIQR)

Where Q1 is the 25th percentile, and Q2 is the 75th percentile.

An additional method to detect outliers involves converting our variables to z-scores and assessing how many values exceed critical limits. In a normal distribution, approximately 5% of data points are expected to have absolute z-scores greater than 1.96 (typically rounded to 2), while 1% surpass 2.58, and none should exceed 3.29 (Field et al., 2012, p.146). Consequently, both the Income and Number_of_trips columns are flagged for outliers as their maximum values exceed these upper thresholds. In the visual representations within this report, any data points surpassing 3.29 will be considered extreme outliers and may be omitted from the plots.

Figure 1

Columns with NA's

```
> # Check for missing values
> colSums(is.na(traveldf))
           CustomerID                   Age         TypeofContact              CityTier            Occupation                Gender
                    0                   226                    25                     0                     0                     0
 NumberOfPersonVisiting     NumberOfFollowups  PreferredPropertyStar         MaritalStatus              Passport PitchSatisfactionScore
                    0                    45                    26                     0                     0                     0
               OwnCar NumberOfChildrenVisiting                Income         NumberOfTrips
                    0                    66                   233                   140
```

Figure 2

Columns with NA's

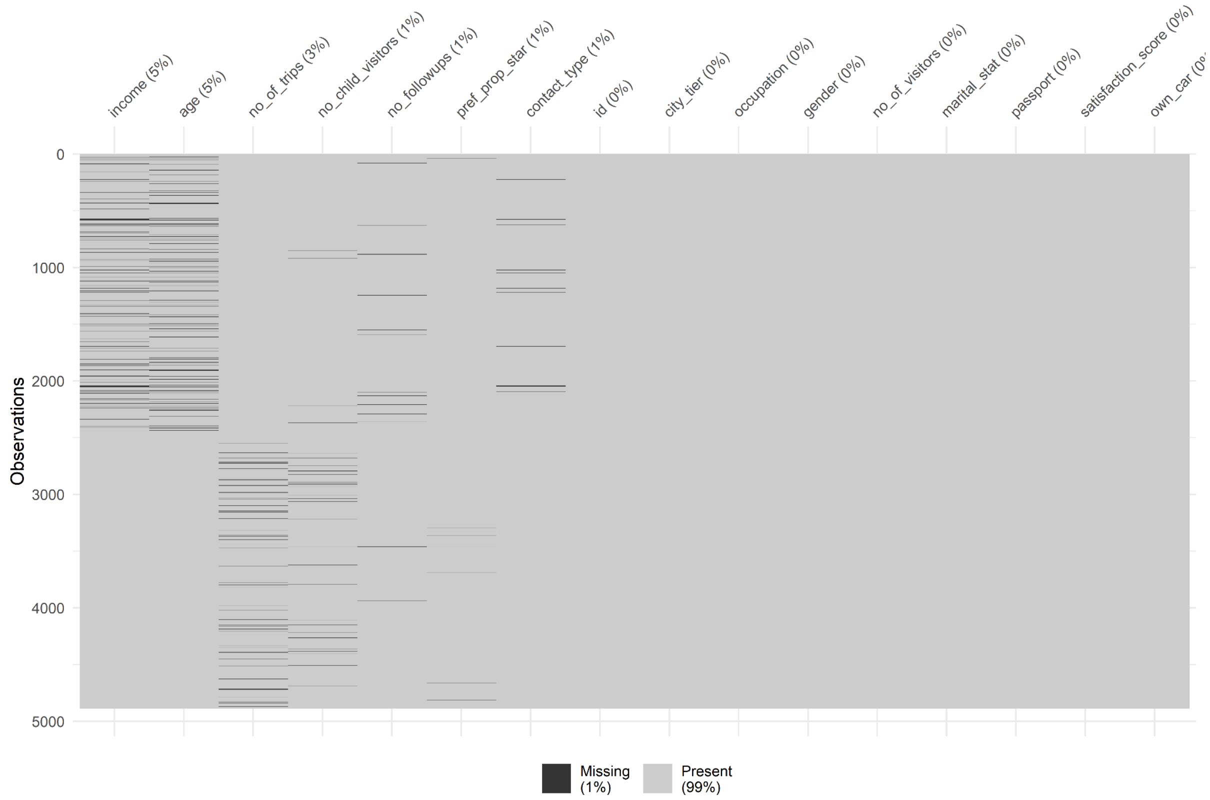


Figure 3

Summary of Numeric Columns

```
> summary(traveldf)
  CustomerID        Age         TypeofContact         CityTier      Occupation
 Min.   :   1   Min.   :18.00   Length:4888        Min.   :1.000   Length:4888
 1st Qu.:1223   1st Qu.:31.00   Class :character   1st Qu.:1.000   Class :character
 Median :2444   Median :36.00   Mode  :character   Median :1.000   Mode  :character
 Mean   :2444   Mean   :37.62                      Mean   :1.654
 3rd Qu.:3666   3rd Qu.:44.00                      3rd Qu.:3.000
 Max.   :4888   Max.   :61.00                      Max.   :3.000
                NA's   :226
    Gender          NumberOfPersonVisiting NumberOfFollowups PreferredPropertyStar
 Length:4888        Min.   :1.000          Min.   :1.000     Min.   :3.000
 Class :character   1st Qu.:2.000          1st Qu.:3.000     1st Qu.:3.000
 Mode  :character   Median :3.000          Median :4.000     Median :3.000
                    Mean   :2.905          Mean   :3.708     Mean   :3.581
                    3rd Qu.:3.000          3rd Qu.:4.000     3rd Qu.:4.000
                    Max.   :5.000          Max.   :6.000     Max.   :5.000
                                           NA's   :45        NA's   :26
 MaritalStatus         Passport      PitchSatisfactionScore     OwnCar
 Length:4888        Min.   :0.0000   Min.   :1.000          Min.   :0.0000
 Class :character   1st Qu.:0.0000   1st Qu.:2.000          1st Qu.:0.0000
 Mode  :character   Median :0.0000   Median :3.000          Median :1.0000
                    Mean   :0.2909   Mean   :3.078          Mean   :0.6203
                    3rd Qu.:1.0000   3rd Qu.:4.000          3rd Qu.:1.0000
                    Max.   :1.0000   Max.   :5.000          Max.   :1.0000

 NumberOfChildrenVisiting     Income      NumberOfTrips
 Min.   :0.000            Min.   : 1000   Min.   : 1.000
 1st Qu.:1.000            1st Qu.:20346   1st Qu.: 2.000
 Median :1.000            Median :22347   Median : 3.000
 Mean   :1.187            Mean   :23620   Mean   : 3.237
 3rd Qu.:2.000            3rd Qu.:25571   3rd Qu.: 4.000
 Max.   :3.000            Max.   :98678   Max.   :22.000
 NA's   :66               NA's   :233     NA's   :140
```

Table 1

Variable Types

Column Name	Variable Type
Age	Discrete
TypeofContact	Nominal
CityTier	Ordinal
Occupation	Nominal
Gender	Nominal
NumberofPersonVisitin	Discrete
NumberofFollowups	Discrete
PreferredPropertyStar	Ordinal
MaritalStatus	Nominal
Passport	Nominal
PitchSatisfactionScore	Ordinal
OwnCar	Nominal
NumberofChildrenVisiting	Discrete
Income	Continuous
NumberofTrips	Discrete

Exploratory Data Analysis (EDA)

Visualisations offer an insightful preliminary glimpse into your data, fostering a deeper understanding even before diving into detailed analysis (Field et al., 2012, p.117).

Univariate Plots

Figures 4 to 11 represent histograms displaying the distributions of discrete or continuous numeric variables. The histograms depict near-normal distributions, except for Figures 8, 9, 10, and 11. The presence of outliers causes skewness in Figures 8 and 10. Despite outlier treatment in the 'no_of_trips' variable (Figure 9), it maintains a right skew (mode < median < mean) (LibreTexts, 2023). The near-normal distributions indicate minimal differences between the mean, median, and mode of the numeric columns.

Figures 12 to 20 exhibit bar plots representing categorical variables. To prevent misinterpretation, the y-axis (Frequency) is scaled to match the category with the highest occurrence. Generally, the disparity between the modal category and others surpasses 1000. At times, this suggests an underrepresentation of specific categories. For instance, in Figure 18 showcasing occupations, the "freelance" category registers only 2 occurrences, accounting for less than 1% of the total observations. Such scarcity could adversely impact our model if we aim to predict outcomes for customers within this category.

Figure 4

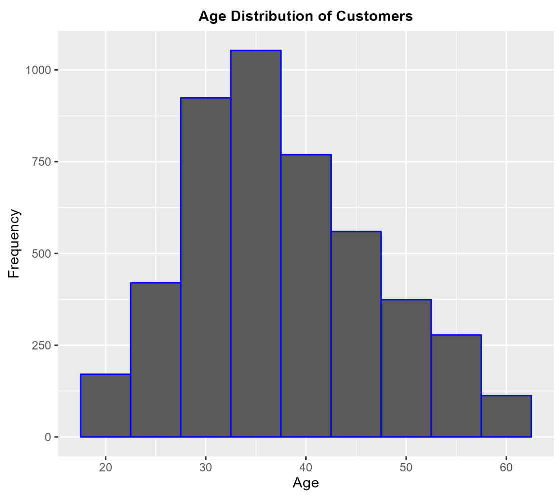


Figure 5

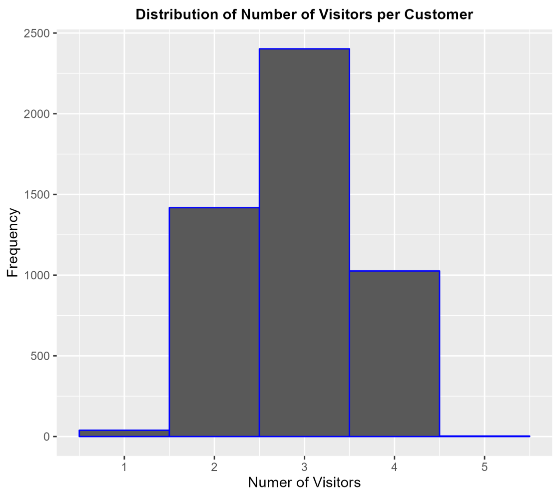


Figure 6

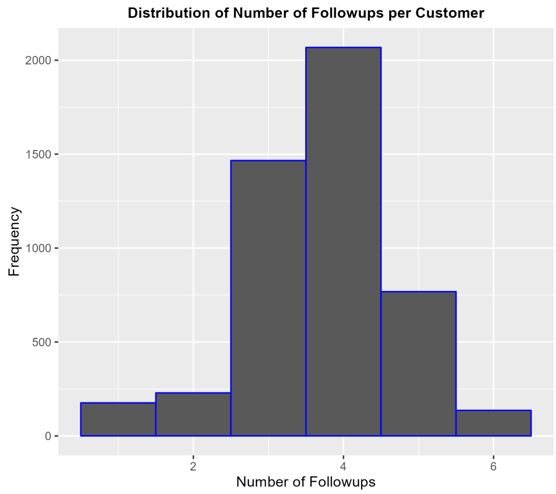


Figure 7

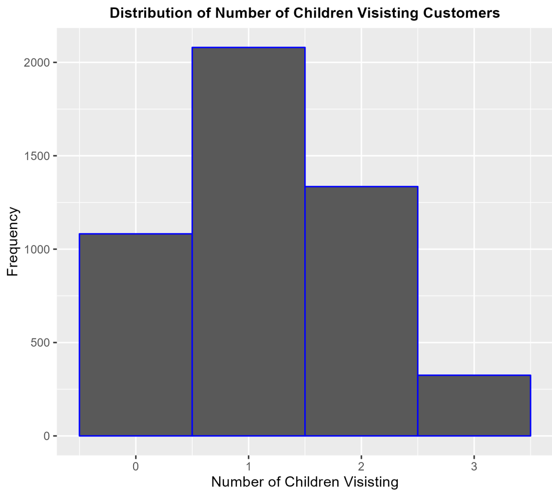


Figure 8

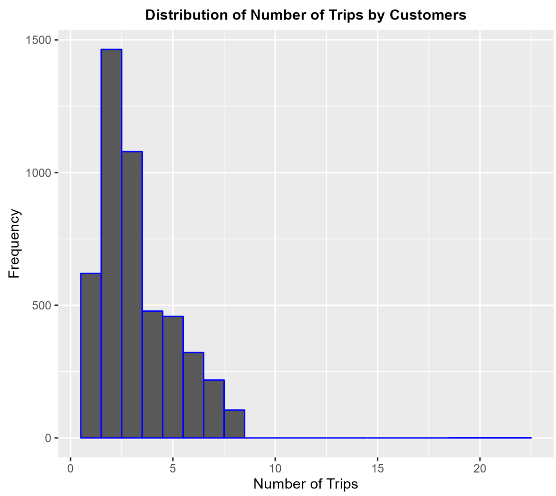


Figure 9

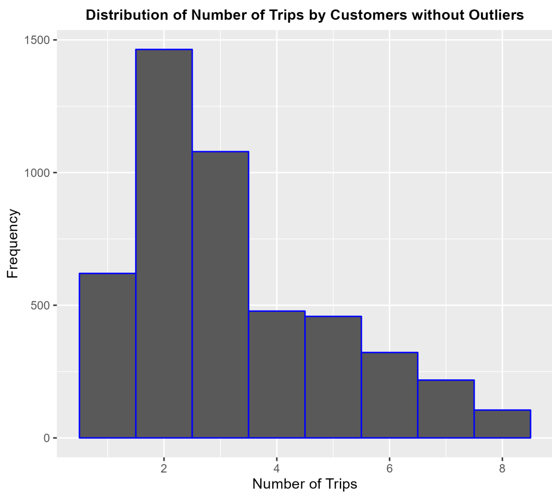


Figure 10

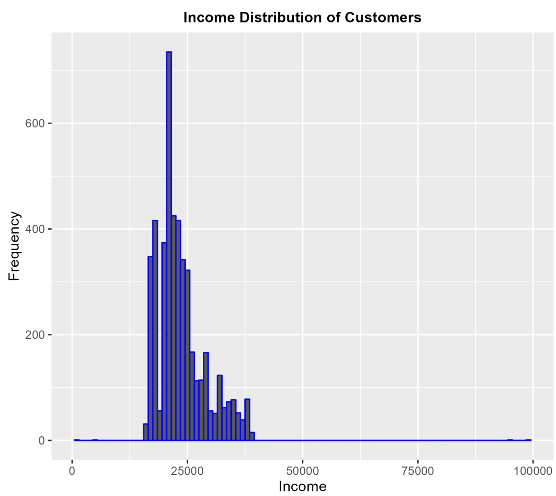


Figure 11

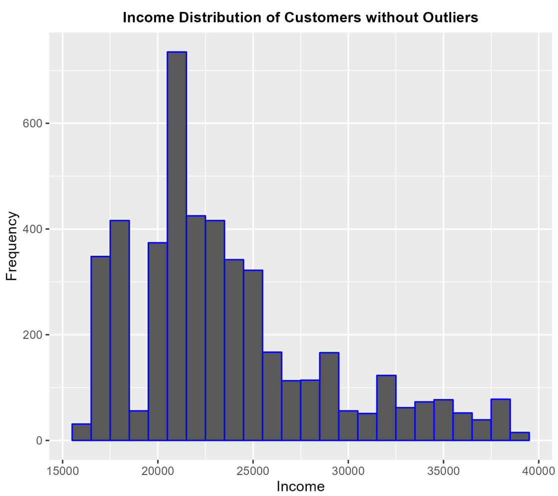


Figure 12

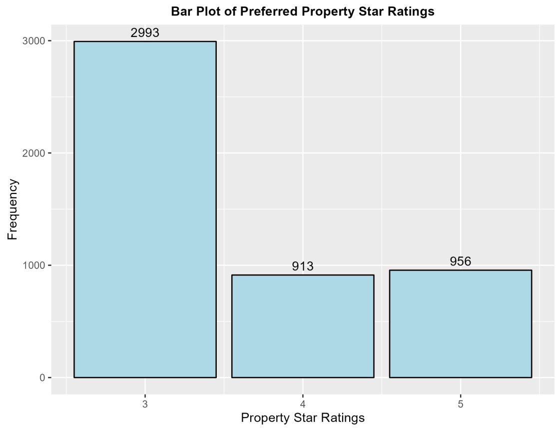


Figure 13

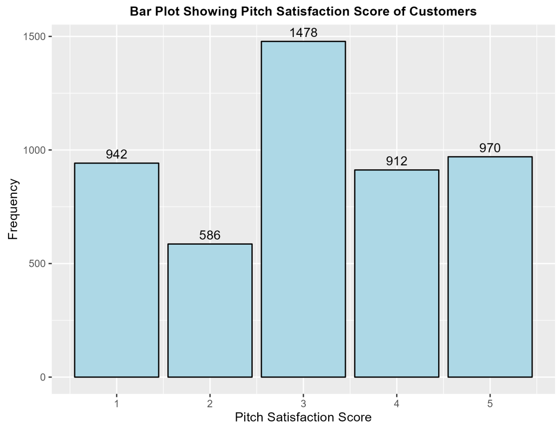


Figure 14

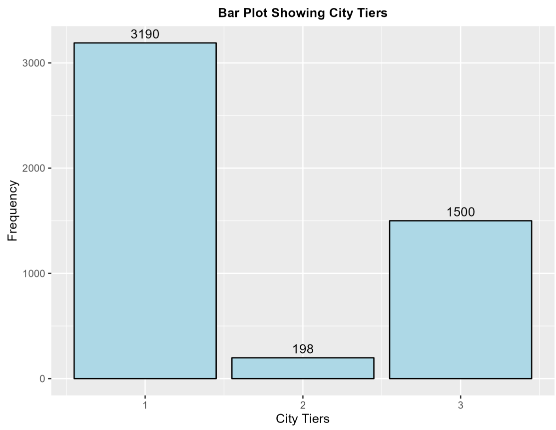


Figure 15

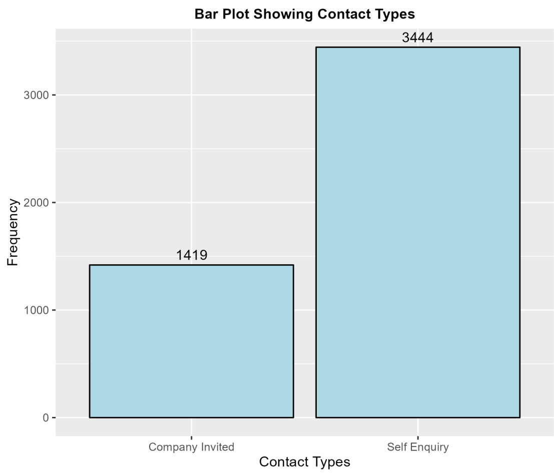


Figure 16

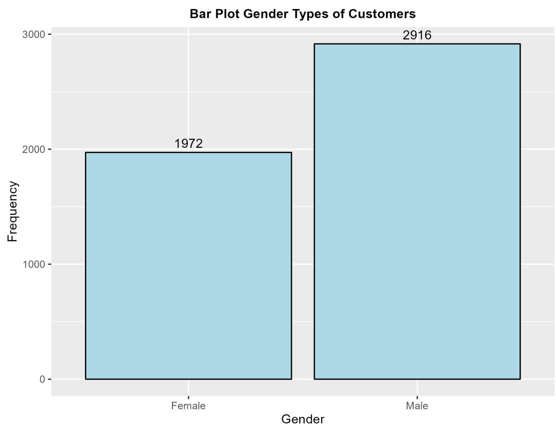


Figure 17

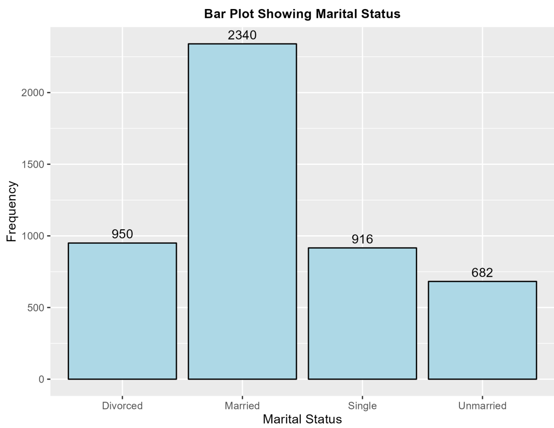


Figure 18

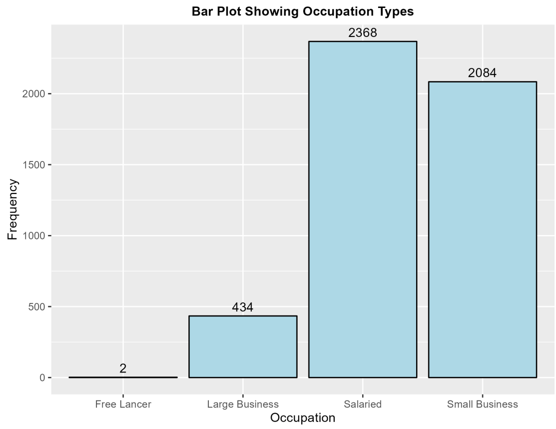


Figure 19

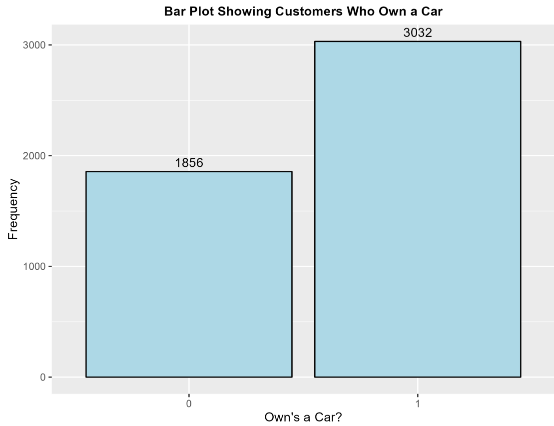
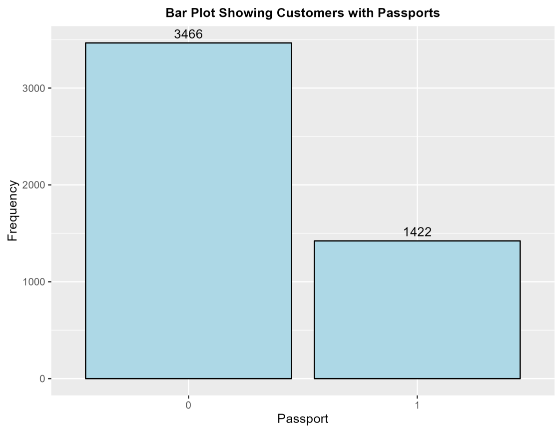


Figure 20



Bivariate Plots

Bivariate plots illustrate the relationship between two variables measured within a single subject sample (Bliwise, nd). Specifically, scatter plots assist in examining trends among numerical variables, allowing us to identify potential linear relationships. Figures 21 to 25 present scatter plots depicting our target variable "no_of_trips" against various numeric columns. Upon reviewing these plots, no distinct relationships appear apparent. However, this absence of observable relationships does not entirely negate the potential correlations between the variables; rather, it indicates a probable lack of strong correlation among them.

The median values across the majority of categorical variable categories appear similar in Figures 26 to 34. However, a few exceptions arise, such as the occupation boxplot where freelancers exhibit a considerably higher average number of trips compared to other occupation categories. Nonetheless, this observation is influenced by the notably low count of observations (2) within this category, as previously noted. Similarly, in Figure 33: “Number of Trips vs city_tier”, a comparable scenario is evident, as depicted in Figure 14. One notable trend observed from this series of plots is that individuals categorised as "single" tend to undertake fewer trips on average than those in other categories, highlighted in Figure 26.

Figure 21

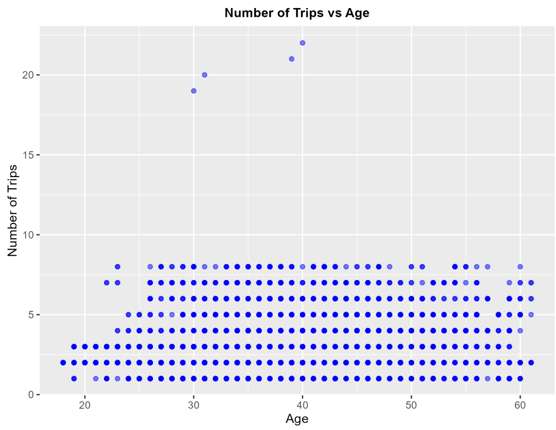


Figure 22

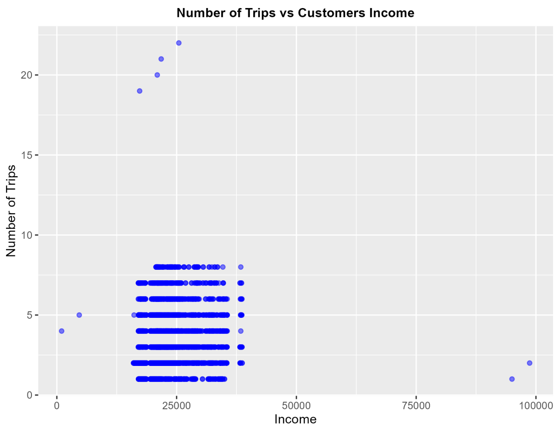


Figure 23

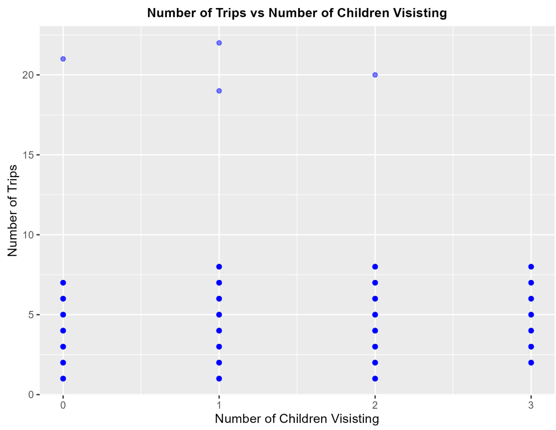


Figure 24

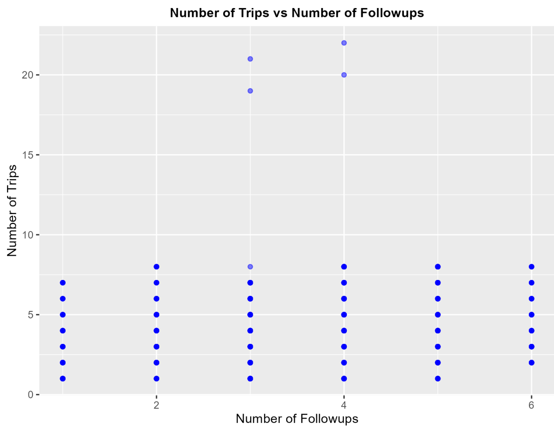


Figure 25

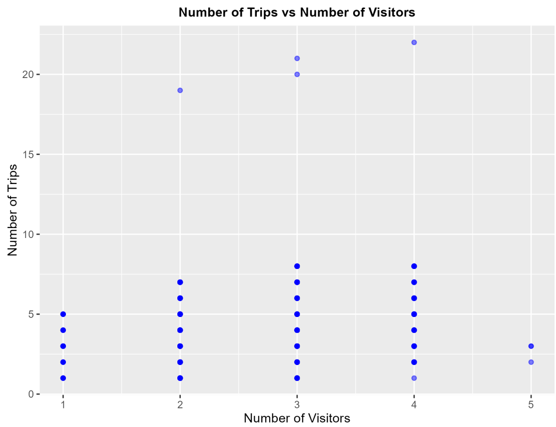


Figure 26

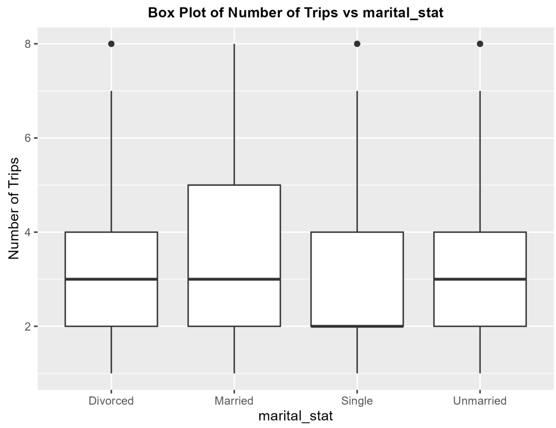


Figure 27

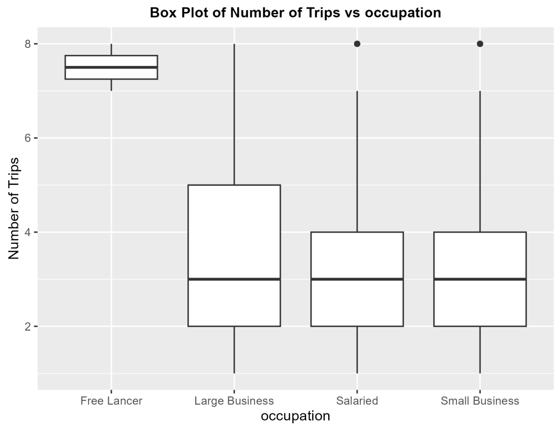


Figure 28

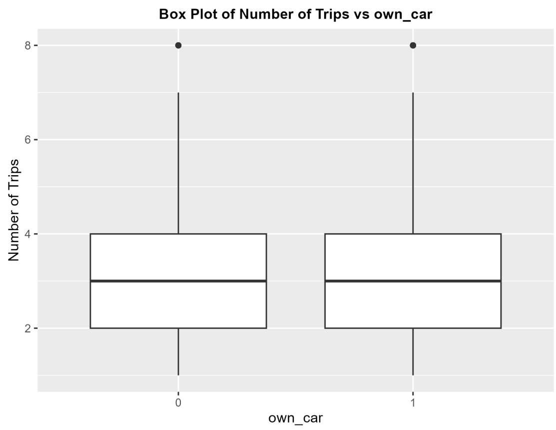


Figure 29

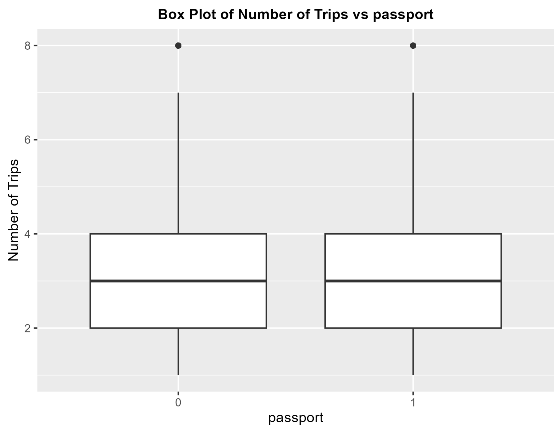


Figure 30

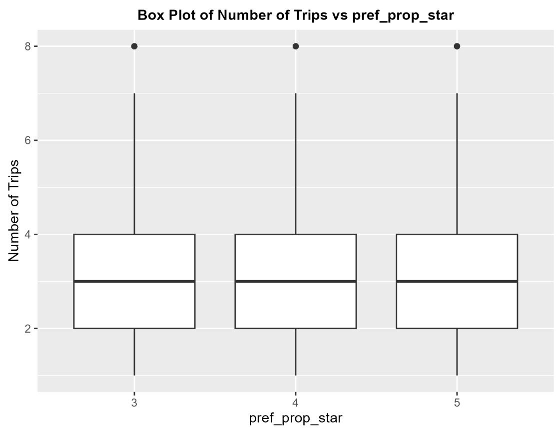


Figure 31

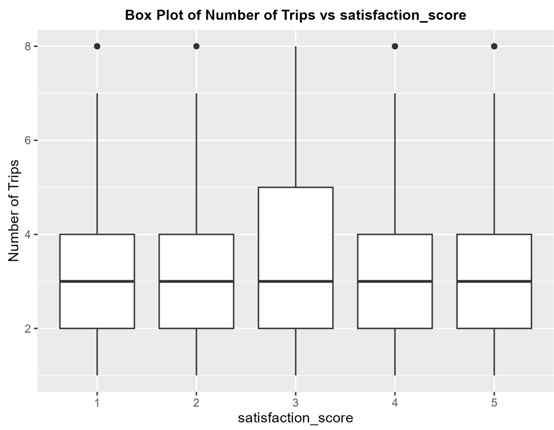


Figure 32

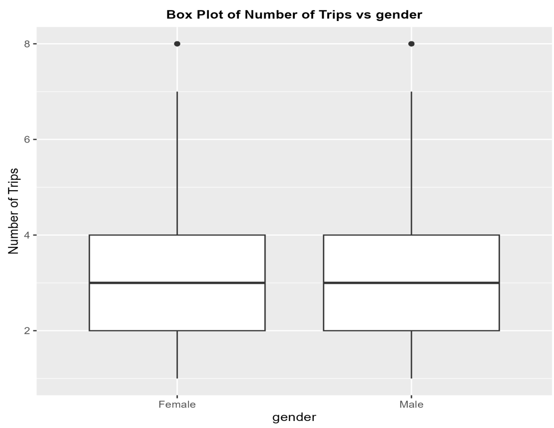


Figure 33

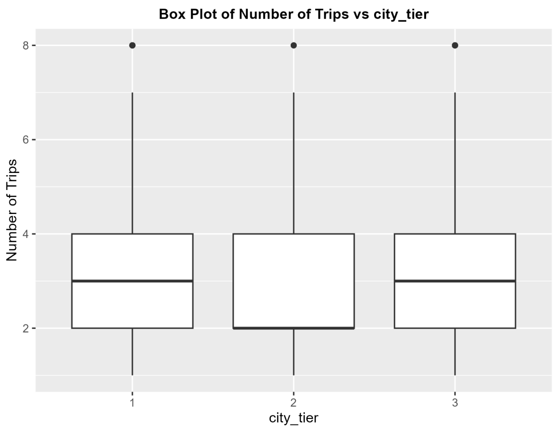
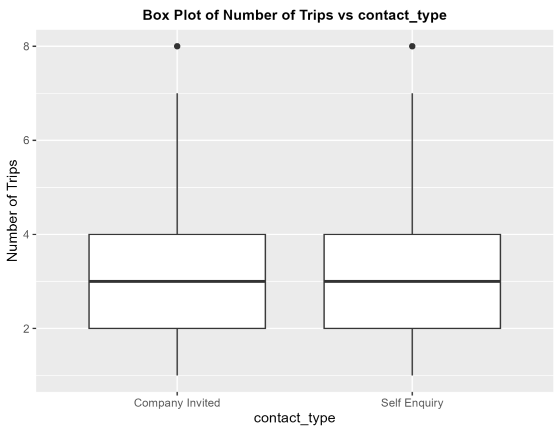


Figure 34



Correlation Analysis

Tables 2 and 3 showcase the correlation between variables and their respective p-values. Kendall's tau-b was chosen primarily for its robustness, considering the violation of parametric assumptions across various segments of our dataset (Field et al., 2012, p. 223). The transformation of character columns into numeric values was executed to conduct this test, following a sequential assignment from 1, exemplified in Figure 35. Notably, correlations and associated p-values presented in red denote statistical significance, indicated by p-values < 0.05. Specifically, variables such as age, number of visitors, follow-ups, children visitors, and income exhibit small yet statistically significant correlations with the "no_of_trips" variable, ranging between 0.13 to 0.20. Conversely, marital status and city tier variables demonstrate relatively weaker correlations compared to others, yet remain statistically significant, potentially contributing to enhancing the model's performance. Tables 2 and 3 were created using the R codes below in Figure 36.

Figure 35

Showing column conversion to numeric values

```
traveldf_corr$contact_type <- as.numeric(factor(traveldf_corr$contact_type ,
   levels = c("Self Enquiry","Company Invited")))
```

Figure 36

Kendall's Correlation Plot Codes

```
#                           CORRELATION PLOT
# Get the names of the columns/variables in your dataframe
vars <- names(traveldf_corr)
# Initialize empty matrices to store correlation coefficients and p-values
cor_matrix <- matrix(NA, ncol = length(vars), nrow = length(vars))
p_value_matrix <- matrix(NA, ncol = length(vars), nrow = length(vars))
# Set column and row names to original column names
dimnames(cor_matrix) <- list(vars, vars)
dimnames(p_value_matrix) <- list(vars, vars)
# Loop through each pair of variables and calculate the correlation and p-value
for (i in 1:length(vars)) {
  for (j in 1:length(vars)) {
    # Compute correlation and test for significance
    test_result <- cor.test(traveldf_corr[[i]], traveldf_corr[[j]], method = "pearson", use = "complete.obs")
    # Store correlation coefficient in the correlation matrix
    cor_matrix[i, j] <- cor(traveldf_corr[[i]], traveldf_corr[[j]], use = "complete.obs")
    # Store p-value in the p-value matrix
    p_value_matrix[i, j] <- test_result$p.value
  }
}
View(cor_matrix)
View(p_value_matrix)
# Export the correlation matrix to a CSV file
write.csv(cor_matrix,
file = "C:/Users/hp/Desktop/For_PJ/MSc Data Analytics/Module 1/MSc_DA_Code_Files/Module1-R/correlation_matrix.csv")
# Export the p-value matrix to a CSV file
write.csv(p_value_matrix,
file = "C:/Users/hp/Desktop/For_PJ/MSc Data Analytics/Module 1/MSc_DA_Code_Files/Module1-R/p_value_matrix.csv")
```

Table 2

Kendall's tau Correlation Plot

	age	contact_type	city_tier	occupation	gender	no_of_visitors	no_followups	pref_prop_star	marital_stat	passport	satisfaction_score	own_car	no_child_visitors	income	no_of_trips
age	1.0000	-0.0261	-0.0156	-0.0109	-0.0393	0.0116	-0.0026	-0.0105	0.0368	-0.0334	0.0185	-0.0487	0.0074	0.4649	0.1849
contact_type	-0.0261	1.0000	0.0096	-0.0360	-0.0005	0.0009	0.0147	-0.0331	0.0080	0.0029	0.0176	0.0034	0.0041	-0.0273	-0.0118
city_tier	-0.0156	0.0096	1.0000	0.1205	-0.0218	-0.0017	0.0237	-0.0092	0.0468	-0.0018	-0.0422	-0.0038	0.0007	0.0518	-0.0297
occupation	-0.0109	-0.0360	0.1205	1.0000	-0.0113	-0.0097	0.0023	0.0335	-0.0097	-0.0025	-0.0117	0.0197	0.0041	-0.0156	0.0160
gender	-0.0393	-0.0005	-0.0218	-0.0113	1.0000	-0.0087	-0.0042	-0.0237	0.0043	0.0370	0.0019	0.0161	0.0211	-0.0349	-0.0025
no_of_visitors	0.0116	0.0009	-0.0017	-0.0097	-0.0087	1.0000	0.3286	0.0339	0.1583	-0.0112	-0.0196	-0.0104	0.6106	0.1951	0.1952
no_followups	-0.0026	0.0147	0.0237	0.0023	-0.0042	0.3286	1.0000	-0.0242	0.0926	-0.0050	0.0041	-0.0121	0.2864	0.1765	0.1395
pref_prop_star	-0.0105	-0.0331	-0.0092	0.0335	-0.0237	0.0339	-0.0242	1.0000	0.0092	-0.0010	-0.0227	-0.0157	0.0358	0.0143	0.0121
marital_stat	0.0368	0.0080	0.0468	-0.0097	0.0043	0.1583	0.0926	0.0092	1.0000	0.0195	-0.0267	-0.0074	0.1292	0.0823	0.0715
passport	-0.0334	0.0029	-0.0018	-0.0025	0.0370	-0.0112	-0.0050	-0.0010	0.0195	1.0000	-0.0029	-0.0223	-0.0203	-0.0025	-0.0129
satisfaction_score	0.0185	0.0176	-0.0422	-0.0117	0.0019	-0.0196	0.0041	-0.0227	-0.0267	-0.0029	1.0000	-0.0688	0.0009	0.0304	-0.0044
own_car	-0.0487	0.0034	-0.0038	0.0197	0.0161	-0.0104	-0.0121	-0.0157	-0.0074	-0.0223	-0.0688	1.0000	-0.0266	-0.0803	0.0118
no_child_visitors	0.0074	0.0041	0.0007	0.0041	0.0211	0.6106	0.2864	0.0358	0.1292	-0.0203	0.0009	-0.0266	1.0000	0.2016	0.1688
income	0.4649	-0.0273	0.0518	-0.0156	-0.0349	0.1951	0.1765	0.0143	0.0823	-0.0025	0.0304	-0.0803	0.2016	1.0000	0.1391
no_of_trips	0.1849	-0.0118	-0.0297	0.0160	-0.0025	0.1952	0.1395	0.0121	0.0715	-0.0129	-0.0044	0.0118	0.1688	0.1391	1.0000

Table 3

Kendall's tau Correlation Plot Corresponding p-values

	age	contact_type	city_tier	occupation	gender	no_of_visitors	no_followups	pref_prop_star	marital_stat	passport	satisfaction_score	own_car	no_child_visitors	income	no_of_trips
age	0.0000	0.0759	0.2861	0.4579	0.0073	0.4276	0.8610	0.4759	0.0119	0.0226	0.2064	0.0009	0.6174	0.0000	0.0000
contact_type	0.0759	0.0000	0.5020	0.0121	0.9745	0.9503	0.3074	0.0211	0.5792	0.8389	0.2205	0.8131	0.7778	0.0630	0.4167
city_tier	0.2861	0.5020	0.0000	0.0000	0.1283	0.9070	0.0998	0.5229	0.0011	0.9002	0.0032	0.7896	0.9628	0.0004	0.0407
occupation	0.4579	0.0121	0.0000	0.0000	0.4284	0.4965	0.8756	0.0195	0.4994	0.8636	0.4121	0.1687	0.7761	0.2875	0.2708
gender	0.0073	0.9745	0.1283	0.4284	0.0000	0.5412	0.7704	0.0980	0.7636	0.0097	0.8962	0.2593	0.1426	0.0174	0.8606
no_of_visitors	0.4276	0.9503	0.9070	0.4965	0.5412	0.0000	0.0000	0.0182	0.0000	0.4347	0.1711	0.4689	0.0000	0.0000	0.0000
no_followups	0.8610	0.3074	0.0998	0.8756	0.7704	0.0000	0.0000	0.0934	0.0000	0.7295	0.7779	0.3994	0.0000	0.0000	0.0000
pref_prop_star	0.4759	0.0211	0.5229	0.0195	0.0980	0.0182	0.0934	0.0000	0.5230	0.9422	0.1135	0.2724	0.0132	0.3311	0.4052
marital_stat	0.0119	0.5792	0.0011	0.4994	0.7636	0.0000	0.0000	0.5230	0.0000	0.1725	0.0621	0.6055	0.0000	0.0000	0.0000
passport	0.0226	0.8389	0.9002	0.8636	0.0097	0.4347	0.7295	0.9422	0.1725	0.0000	0.8380	0.1185	0.1594	0.8622	0.3724
satisfaction_score	0.2064	0.2205	0.0032	0.4121	0.8962	0.1711	0.7779	0.1135	0.0621	0.8380	0.0000	0.0000	0.9514	0.0379	0.7629
own_car	0.0009	0.8131	0.7896	0.1687	0.2593	0.4689	0.3994	0.2724	0.6055	0.1185	0.0000	0.0000	0.0650	0.0000	0.4153
no_child_visitors	0.6174	0.7778	0.9628	0.7761	0.1426	0.0000	0.0000	0.0132	0.0000	0.1594	0.9514	0.0650	0.0000	0.0000	0.0000
income	0.0000	0.0630	0.0004	0.2875	0.0174	0.0000	0.0000	0.3311	0.0000	0.8622	0.0379	0.0000	0.0000	0.0000	0.0000
no_of_trips	0.0000	0.4167	0.0407	0.2708	0.8606	0.0000	0.0000	0.4052	0.0000	0.3724	0.7629	0.4153	0.0000	0.0000	0.0000

Model Development

The seven aforementioned variables in the correlation section of this report will be incorporated into the initial model due to their observed associations with the target variable. However, it's crucial to exercise caution regarding the number of predictor variables in the model. As a fundamental guideline, a smaller count of predictors tends to yield better outcomes (Field, et al., 2012, p. 266).

In order to construct our model, it's imperative to handle missing values, as their presence may lead to errors when attempting to estimate a model using data frames (Field, et al., 2012, p. 257). The kNN (k-Nearest Neighbors) method will be employed for addressing these missing values within the dataset. This technique involves imputing missing values of an attribute by utilising a specified number of attributes that closely resemble the attribute with missing values (Mekala, 2018). To execute this process, the kNN function from the VIM package in RStudio will be utilised, as illustrated in Figure 37(Oleszak, nd). The code illustrated in Figure 37 was formulated based on a tutorial from DataCamp taught by Michal Oleszak.

The initial model, labelled "tripping_mlrm" (Figure 38), indicates that marital status factors with Pr(>|t|) > 0.01 do not notably contribute to the predictions. However, for this model, variables such as no_of_visitors (t(4878) = 0.3019, p < .001), age (t(4878) = 0.0342, p < .001), no_child_visitors (t(4878) = 0.1381, p < .001), no_followups (t(4878) = 0.1413, p < .001), and city_tier (t(4878) = -0.0601, p < .001) are all deemed significant predictors of the number_of_trip variable. With probabilities (Pr(>|t|)) less than .001 for these variables except city_tier which is less than .05, we infer that the probability of their t-values (or larger) occurring if the values of b in the population were 0 is less than .05. Hence, these five predictor variables significantly contribute to predicting no_of_trips, indicating the potential increase in trips for a 1-unit change in these predictors. For instance, for every 1-unit rise in age, the trips increase by 0.0342, as demonstrated by the "Estimate" (b value) in Figure 38. Furthermore, the intercept (b_o) (t(4878) = 0.4760, p < .05) is also deemed significant. The

F-statistic with a p-value < 2.2e-16 indicates that our model significantly outperforms the mean model in predicting no_of_trips. However, our model only explains approximately 7.7% (0.0773x100) of the variation in no_of_trips, as indicated by the multiple R^2 (Figure 38) (Field et al, 2012, p.258-260).

The equation for predicting our target variable based on these values will be given as:

no_of_trips = 0.4760 + 0.3019(no_of_visitors) + 0.0342(age) + 0.1381(no_child_visitors) + 0.1413(no_followups) -0.0601(city_tier)

There's a need to check for outliers and influential cases using standardised residuals and Cook's distance (Figure 39). Standardised residuals represent model prediction errors for specific observations, while Cook's distance gauges an observation's overall influence on the model (Field et al., 2012, p.268-269). The VIF values showed no threat of multicollinearity (Myers, 1990). Subsequently, after excluding observations with substantial residuals and utilising the all-subset method to determine the best model, our refined model improved the R^2 to 8.2% (Figure 40). This enhanced model serves as the basis for predictions in the subsequent section of this report.

Figure 37

Dealing with Missing Values

```
traveldf_no_na <- kNN(traveldf_model, k=5,
                      variable = c("contact_type","pref_prop_star","no_followups",
                       "no_child_visitors","no_of_trips","age","income"))
```

Figure 38

Initial Model "tripping_mlrm"

```
> tripping_mlrm <- lm(no_of_trips ~ no_of_visitors + age + no_child_visitors +
+                      income + no_followups + marital_stat + city_tier,
+                      data = traveldf_no_na)
> # now lets evaluate the initial model
> summary(tripping_mlrm)

Call:
lm(formula = no_of_trips ~ no_of_visitors + age + no_child_visitors +
    income + no_followups + marital_stat + city_tier, data = traveldf_no_na)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1404 -1.2305 -0.4250  0.9424 18.2504

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            4.760e-01  1.879e-01   2.533 0.011344 *
no_of_visitors         3.019e-01  4.516e-02   6.687 2.54e-11 ***
age                    3.424e-02  3.122e-03  10.968  < 2e-16 ***
no_child_visitors      1.381e-01  3.755e-02   3.678 0.000237 ***
income                -6.003e-07  5.569e-06  -0.108 0.914171
no_followups           1.413e-01  2.719e-02   5.196 2.12e-07 ***
marital_statMarried    6.685e-02  6.782e-02   0.986 0.324342
marital_statSingle    -8.781e-02  8.263e-02  -1.063 0.287975
marital_statUnmarried  6.819e-02  8.919e-02   0.765 0.444576
city_tier             -6.012e-02  2.765e-02  -2.175 0.029711 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.761 on 4878 degrees of freedom
Multiple R-squared:  0.0773,	Adjusted R-squared:  0.0756
F-statistic: 45.41 on 9 and 4878 DF,  p-value: < 2.2e-16
```

Figure 39

VIF code snippet

```
> # Calculate the percentage of values greater than 3 in standardres
> percentage_gt_3 <- mean(traveldf_lrmres$standardres > 3) * 100
> percentage_gt_3
[1] 0.08183306
> #Now let's check for the cook distance
> traveldf_lrmres$cook <- cooks.distance(tripping_mlrm)
> pct_cooks_gt_1 <- mean(traveldf_lrmres$cook > 1) * 100
> pct_cooks_gt_1
[1] 0
> # Calculate VIF for the model
> vif_result <- car::vif(tripping_mlrm)
> vif_result
                      GVIF Df GVIF^(1/(2*Df))
no_of_visitors    1.688409  1        1.299388
age               1.297687  1        1.139161
no_child_visitors 1.625242  1        1.274850
income            1.390000  1        1.178982
no_followups      1.164312  1        1.079033
marital_stat      1.067756  3        1.010986
city_tier         1.011894  1        1.005929
```

Figure 40

Final model tripping_mlrm3 code snippet

```
> tripping_mlrm3 <- lm(no_of_trips ~ no_of_visitors + age + no_child_visitors +
+                       no_followups,
+                    data = traveldf_lrmres_ls3)
> summary(tripping_mlrm3)

Call:
lm(formula = no_of_trips ~ no_of_visitors + age + no_child_visitors +
    no_followups, data = traveldf_lrmres_ls3)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1764 -1.2098 -0.4172  0.9574  5.1382

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.324064   0.157371   2.059   0.0395 *
no_of_visitors    0.301168   0.042954   7.011 2.68e-12 ***
age               0.035057   0.002632  13.320  < 2e-16 ***
no_child_visitors 0.149070   0.035859   4.157 3.28e-05 ***
no_followups      0.143418   0.025778   5.564 2.78e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.691 on 4879 degrees of freedom
Multiple R-squared:  0.08224,	Adjusted R-squared:  0.08149
F-statistic: 109.3 on 4 and 4879 DF,  p-value: < 2.2e-16
```

Predicting Test Cases

The model generates trip predictions in decimal points, yet since "no_of_trips" is a discrete value, the forecasted values are rounded to the nearest whole number, reflected in the "rounded_(y)" column (Table 4). The code used to derive predictions from the "tripping_mlrm3" model (Figure 40) is outlined in Figure (41). It's crucial to emphasise that these forecasts stem from a model elucidating merely around 8% of the variance in our target variable. Hence, approximately 92% of the variation remains unexplained, likely attributed to other influencing factors.

Figure 41

Model Prediction code snippet

```r
# Extract the columns used in the model
cols_used <- c("no_of_visitors", "age", "no_child_visitors", "no_followups")

# Make predictions on the test dataset using the specified columns
predictions <- predict(tripping_mlrm3, newdata = traveldf_test_colsrenamed[, cols_used])

# Append the predictions to the test dataset
traveldf_test_colsrenamed$predicted_no_of_trips <- predictions
traveldf_test_colsrenamed$Rpredicted_no_of_trips <- round(predictions,0)
View(traveldf_test_colsrenamed)
```

Table 4

Model Predictions

id	age	contact_type	city_tie	occupation	gender	no_of_visitc	no_followup	pref_prop_st	marital_st	passpo	satisfaction_sc	own_ca	no_child_visitc	income	**predicted_no_of_trip**	**rounded_(y**
1	42	Self Enquiry	2	Salaried	Female	2	3	5	Divorced	0	3	0	1	31799	**2.978124867**	**3**
2	30	Self Enquiry	3	Small Busine	Female	3	5	3	Married	1	1	1	2	21478	**3.294512023**	**3**
3	28	Company Invit	1	Salaried	Male	4	5	3	Single	1	3	1	2	21212	**3.525565866**	**4**
4	54	Self Enquiry	1	Salaried	Male	3	4	3	Single	0	5	1	1	21128	**3.843396962**	**4**
5	39	Company Invit	2	Salaried	Male	4	4	4	Married	0	3	1	3	21270	**3.916846709**	**4**

Conclusion

The report outlines findings regarding variable distributions, their interactions, and correlations with the target variable. Among the six numeric variables, four exhibit normal distributions, while the remaining two demonstrate a right-skewed nature. Scatter plots showcasing numeric variables' associations with the target variable didn't reveal linear relationships. Similarly, most categorical variables, as indicated by box plots, exhibited similar median values per category. The correlation analysis showed that the age, no_of_visitors, no_of_followups, no_child_ visitors and income have small significant correlation with the target variable. However, upon further model development steps the income variable proved to be irrelevant in predicting our target variable, and thus wasn't included in the final model. The ultimate model, developed using an all-subset approach, achieved an R^2 of 0.08224, accounting for approximately 8% of the variance in the target variable. With no strong or medium correlations between the 14 predictor variables and the target, there's limited expectation for significant model enhancement. Initial continuous predictions were rounded to the nearest whole numbers due to the discrete nature of the target variable. Future research may explore additional variables to enhance the model and better account for the unexplained 92% variation in customer trip numbers.

References

Bliwise, N.G. (n.d.). Bivariate Plots. Emory College of Arts and Scince https://psychology.emory.edu/clinical/bliwise/Tutorials

Buxton, P., & Tabor, P. (2003, September). Outlier detection for DPPM reduction. In International Test Conference, 2003. Proceedings. ITC 2003. (pp. 818-818). IEEE Computer Society.

Field, A., Miles, J., & Field, Z. (2012). Discovering statistics using R. Sage publications.

Mekala, H. (2018, June). Dealing with Missing Data using R. Medium. Dealing with Missing Data using R | by Harshitha Mekala | Coinmonks | Medium

Myers, J. (1990). Variance inflation factor analysis: Interpretation and understanding.

Oleszak, M. (n.d.) k-Nearest-Neighbors imputation. DataCamp. k-Nearest-Neighbors imputation | R