# Factors contributing to online food ordering activities of customers

### Executive Summary

The paper uses the data about online food delivery in the region of Bangalore, India to analyze factors that can affect the decision to order online food of people in this region. Using some techniques in multivariate analysis, this paper has founded that preference of self-cooking, waiting time, level of affordability, age, and bad past experience have a statistically significant relationship with the online food purchasing decision. This result would be critical for restaurants and food services companies to have appropriate measures and activities to treat their customers and gain more market shares in the online food market, which has grown significantly during COVID 19 pandemic.

The paper is structured as below: part 1 explains the motivation and research questions of this paper. Section 2 conducts some preliminary descriptive analysis of the data set. And then section 3 uses multivariate analysis techniques including checking multivariate normality assumption, logistic regression, multiple linear regression, and principal component analysis to answer the question.

## 1 Research Questions

It's no surprise that food distribution is so common these days, with industry on the increase and a greater appetite for convenience. Food distribution has become a significant trend among customers in all demographics, according to recent studies. According to Frost and Sullivan, the food delivery service industry provided $82 billion in gross sales in 2018, and this figure is expected to more than double by 2025 (Frost and Sullivan, 2019). Today, the COVID-19 pandemic necessitates distribution for foods, and 65 percent of restaurants state that by delivering delivery, they were able to maximize revenues during COVID-19. The COVID-19 lockout forced the closure of restaurants and dining rooms around the world, resulting in significant sales losses for companies in this sector. As the pandemic continues, many of them have taken to food distribution systems as a last resort to stay alive.

Although foot traffic used to be the main driver of restaurant visits, digital now reigns supreme. The majority of consumers use the Internet to locate new restaurants; they often perform online research before deciding on a restaurant. Consumers usually get their restaurant ideas from a mix of search engines, online feedback from consumers and the public, and the restaurant's website and menu. They also use Instagram, Facebook, and other social media platforms to decide when and when to eat based on their diet tastes and financial constraints.

India isn't an outlier in this regard. Restaurants and related services were heavily impacted after the outbreak of COVID19, causing the Indian government to classify food and other related services as critical services. As a result, hotels, restaurants, and food distribution services will now begin operations because they are relied upon by at least 20% of the Indian population, including students, paying visitors, and young professionals (Statista, 2021). The COVID19 pandemic has introduced a new challenge to the food distribution industry and could affect online food delivery services. Restaurants and associated providers, mostly fast-food restaurants, are able to provide food. Customers, on the other hand, remain wary of placing orders during the pandemic, despite the fact that much online food delivery has required their distribution partners to wear personal protective equipment and encouraged customers to pay online to ensure contactless delivery. The health of the people who supply the food and the sanitary conditions of the restaurants are two major factors in the decrease in online food delivery. Present consumers have been prompted to rethink their prospective buying choices as a result of these problems.

Overall, it is important to investigate the factors that influence customers' online food ordering practices in India so that restaurants can take adequate steps to boost customer loyalty and increase online orders during the COVID 19 outbreak. As a result, data on online food ordering in Bangalore will be analyzed to see whether there is a connection between online food ordering and other variables such as age, distribution, food quality, discount, or maximum wait time.

## 2 Descriptive statistics

The data about the preference of online food delivery in the region of Bangalore, India is collected from Kaggle (Ben Roshan. Online Food Delivery Preferences – Bangalore region. Retrieved from: https://www.kaggle.com/benroshan/online-food-delivery-preferencesbangalore-region).

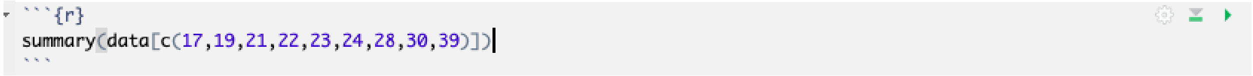
The data consists of 55 variables, including respondents' demographics (age, gender, marital status, occupation, income, education, family size), purchasing decision, and general opinions, preferences, or experiences about online food ordering activities (such as ease and convenient, unaffordable, offers and discount). The number of observations is 388.

First, the demographics of respondents are analyzed and the summary statistics of these variables are provided as below

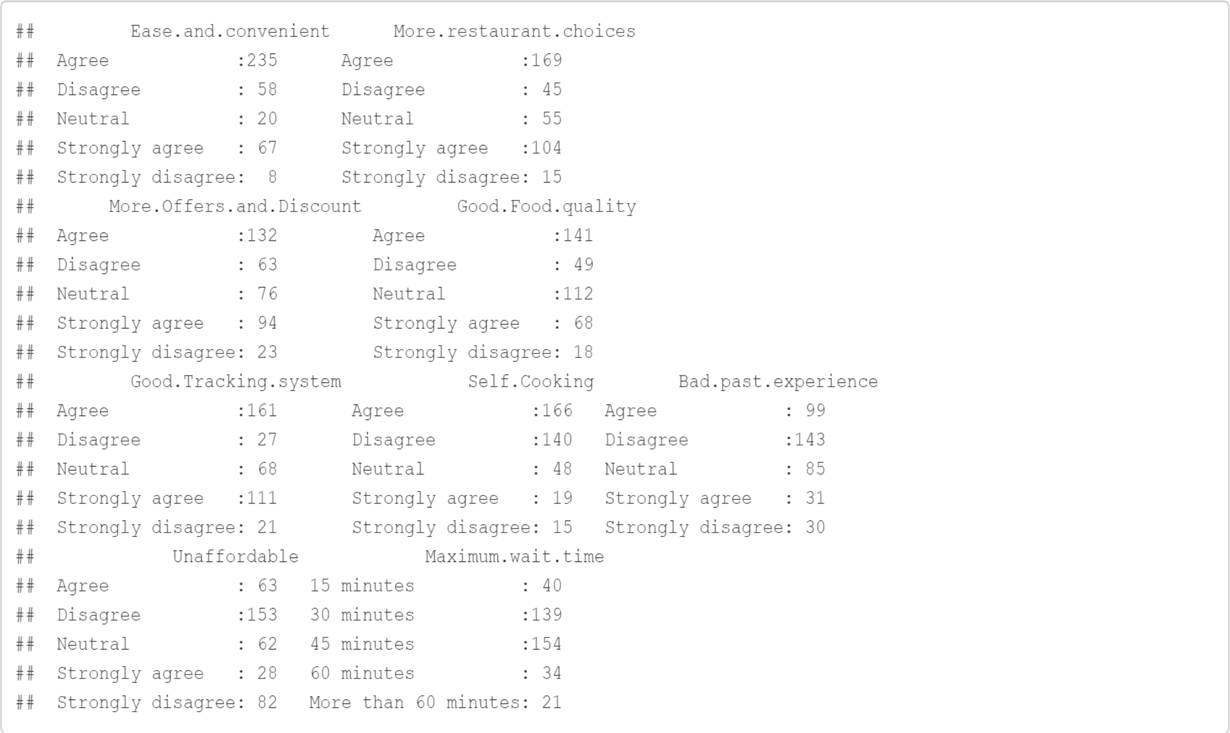
```
summary(data[1:6])
```

```
##       Age           Gender             Marital.Status           Occupation
##  Min.   :18.00   Female:166   Married           :108    Employee      :118
##  1st Qu.:23.00   Male  :222   Prefer not to say: 12    House wife    :  9
##  Median :24.00                Single            :268    Self Employeed: 54
##  Mean   :24.63                                          Student       :207
##  3rd Qu.:26.00
##  Max.   :33.00
##          Monthly.Income Educational.Qualifications
##  10001 to 25000 : 45    Graduate     :177
##  25001 to 50000 : 69    Ph.D         : 23
##  Below Rs.10000 : 25    Post Graduate:174
##  More than 50000: 62    School       : 12
##  No Income      :187    Uneducated   :  2
##
```

Thus, it is clear that respondents are at young ages, as the maximum age of these people is 33, and the average age of respondents is 24.63, indicating that frequent users of online delivery services are normally young people. The main occupations of answerers are a student with 207 over 388 (about 53%), with the majority of them have a low level of income. In addition, almost all repliers at least have a school degree.



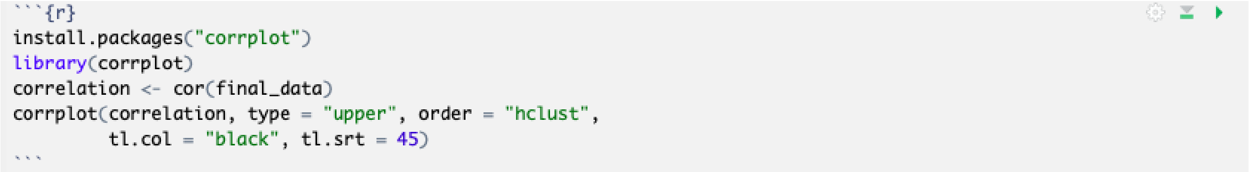
~~~
```{r}
summary(data[c(17,19,21,22,23,24,28,30,39)])
```
~~~



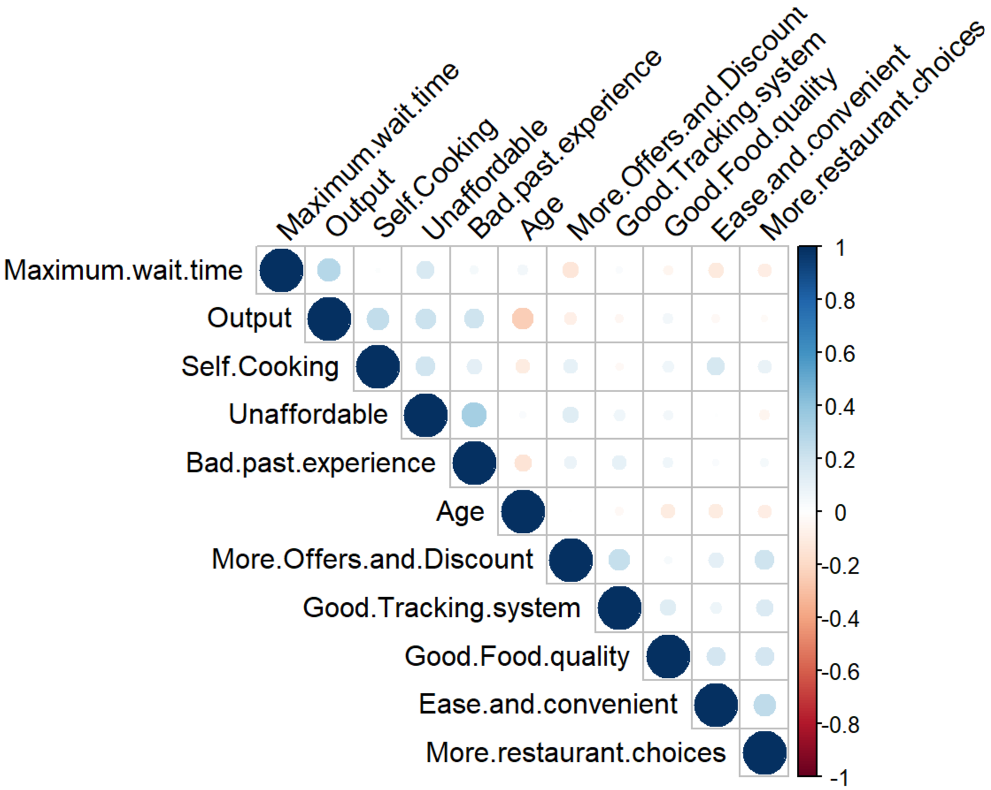
```
##         Ease.and.convenient      More.restaurant.choices
##  Agree            :235      Agree            :169
##  Disagree         : 58      Disagree         : 45
##  Neutral          : 20      Neutral          : 55
##  Strongly agree   : 67      Strongly agree   :104
##  Strongly disagree:  8      Strongly disagree: 15
##       More.Offers.and.Discount         Good.Food.quality
##  Agree            :132         Agree            :141
##  Disagree         : 63         Disagree         : 49
##  Neutral          : 76         Neutral          :112
##  Strongly agree   : 94         Strongly agree   : 68
##  Strongly disagree: 23         Strongly disagree: 18
##         Good.Tracking.system            Self.Cooking        Bad.past.experience
##  Agree            :161       Agree            :166   Agree            : 99
##  Disagree         : 27       Disagree         :140   Disagree         :143
##  Neutral          : 68       Neutral          : 48   Neutral          : 85
##  Strongly agree   :111       Strongly agree   : 19   Strongly agree   : 31
##  Strongly disagree: 21       Strongly disagree: 15   Strongly disagree: 30
##             Unaffordable            Maximum.wait.time
##  Agree            : 63   15 minutes          : 40
##  Disagree         :153   30 minutes          :139
##  Neutral          : 62   45 minutes          :154
##  Strongly agree   : 28   60 minutes          : 34
##  Strongly disagree: 82   More than 60 minutes: 21
```

Next, move to the preferences and experiences of respondents to online food delivery, we can see the ease and convenience are one important factor of this services, as there are 235 over 388 agree with this advantages of online food. Other features of online food services such as more restaurant choices, offers and discounts, food quality, and tracking system also receive much agreement from repliers. Meanwhile, the majority of respondents do not have bad past experiences or find that online food unaffordable.

The correlation matrix is then constructed to find any strong relationships between variables. The correlation graph shows that there is not a strong linear correlation between features, as the biggest correlation is only about 0.33 (between Unaffordable and Bad.past.experience). The online food purchase decision (output variable) seems to have a positive relationship with Self. Cooking, Unaffordable and Bad.past.experience, and negative relationship with Age.



~~~
```{r}
install.packages("corrplot")
library(corrplot)
correlation <- cor(final_data)
corrplot(correlation, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```
~~~



# 3 Multivariate data analysis

We will use methods of Regression and Principal Component Analysis to analyze the impact of factors on the online food delivery preferences of customers. Based on the information and the quality of the data set, the following factors are (1) Ease and convenience, (2) Self Cooking, (3) Maximum wait time, (4) More restaurant choices, (5) Unaffordable, (6 ) More Offers and Discount, (7) Good Food quality, (8) Good Tracking system, (9) Age, (10) Bad past experience will be included in the evaluation model.

## *3.1 Check multivariate normality assumption*

Actually, some statistical methods require a robust assumption about normality. We may construct a Q-Q plot for displaying a distribution, or carry out a systematic statistic test like an Anderson Darling Test or a Jarque-Bera Test if we want to try out whether or not a single variable is normally distributed. However, when approaching multivariate regression, we need to test normality assumption in group factors (Székely and Rizzo, 2005).

Sometimes it depends on the magnitude of the violation to the effect of an inference violation upon the normality test outcome. Any minor breaches can have no real impact on the study, while others may unnecessarily or uninterpretable make the normality test outcome.

The Henze-Zirkler (HZ) test is recommended by many statisticians for the test of multivariate normality. Researchers also tend to use informative and easy-to-understand assessments in many functional applications. Therefore, we would use the Henze-Zirkler Multivariate Normality Test with hypotheses: $H_0$: follow multivariate normal distribution and $H_1$: do not follow a multivariate normal distribution

```
HZ.test(final_data)
```

```
## [1] 0.000000 2.075306
```

Results show that the dataset may have problems of multivariate normality violation. Therefore, the results from the Multiple Regression Analysis methods may not be robust. We need to apply other methods such as Logistic Regression or Principal Component Analysis to provide an overview of the factors affecting customer decisions.

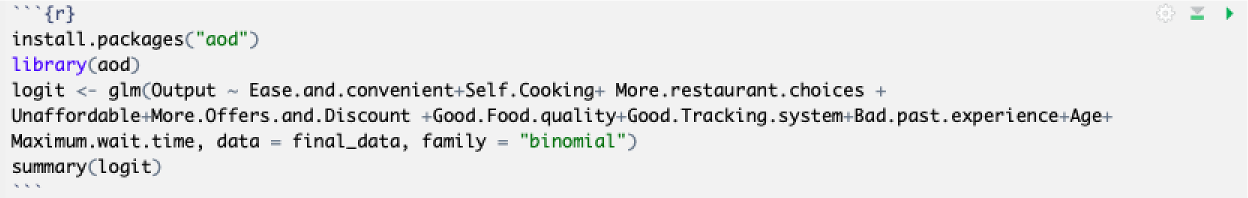
## 3.2 Logistic Regression

We like to characterize our comments as a “select” or “not” user from the website while dealing with our data that builds in a binary distinction. The chance of belonging to someone would be tried by a logistic regression model. The regression of logistics constitutes mostly an extension of linear regression, just [0, 1] is the expected result value. The model would define relations between our target characteristic and our remaining characteristics to calculate probabilistically to determine which class should be a consumer (Kleinbaum et., 2002).

When the dependent variable is dichotomous, logistic regression is the only regression analysis to use (binary). The logistic regression, like all regression analyses, is a statistical analysis. To characterize data and illustrate the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables, logistic regression is used. It is a categorical answer vector that is used in a classification algorithm. The goal of Logistic Regression is to discover a connection between features and the likelihood of a specific outcome. When the amount of hours spent studying is given as a function in predicting whether a student passes or fails an exam, the answer variable has two values: pass and fail.

Binomial Logistic Regression is a type of problem in which the answer variable has two values: 0 and 1, or pass and fail, or true and false. When the answer variable may have three or more potential values, Multinomial Logistic Regression is used.

From the logistic regression output below, we can see that there are six variables that are significant at a 5% significance level. These factors are Ease and convenience, Self Cooking, Unaffordable, Bad past experience, Age, and Maximum wait time.



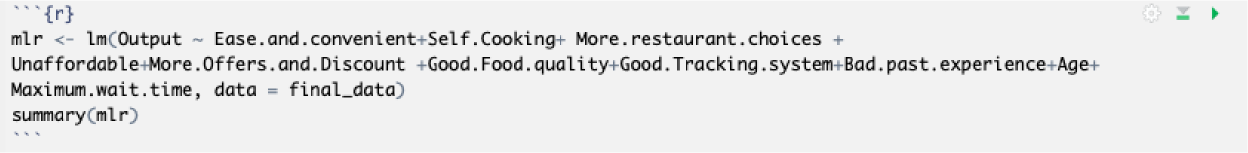
~~~
```{r}
install.packages("aod")
library(aod)
logit <- glm(Output ~ Ease.and.convenient+Self.Cooking+ More.restaurant.choices +
Unaffordable+More.Offers.and.Discount +Good.Food.quality+Good.Tracking.system+Bad.past.experience+Age+
Maximum.wait.time, data = final_data, family = "binomial")
summary(logit)
```
~~~

```
## Call:
## glm(formula = Output ~ Ease.and.convenient + Self.Cooking + More.restaurant.choices +
##     Unaffordable + More.Offers.and.Discount + Good.Food.quality +
##     Good.Tracking.system + Bad.past.experience + Age + Maximum.wait.time,
##     family = "binomial", data = final_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2236   0.0687   0.3697   0.6130   1.5971
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)               1.918379   1.455027   1.318   0.1874
## Ease.and.convenient      -0.274420   0.132068  -2.078   0.0377 *
## Self.Cooking              0.779709   0.192053   4.060 4.91e-05 ***
## More.restaurant.choices  -0.009778   0.114391  -0.085   0.9319
## Unaffordable              0.378365   0.132155   2.863   0.0042 **
## More.Offers.and.Discount -0.087727   0.118278  -0.742   0.4583
## Good.Food.quality         0.156107   0.122255   1.277   0.2016
## Good.Tracking.system     -0.179556   0.108759  -1.651   0.0987 .
## Bad.past.experience       0.359704   0.150384   2.392   0.0168 *
## Age                      -0.210895   0.047585  -4.432 9.34e-06 ***
## Maximum.wait.time         0.981477   0.193205   5.080 3.77e-07 ***
```

## 3.3 Multiple Linear Regression

Multiple linear regression (MLR) is a mathematical method used to forecast the outcomes of a response variable by using many explanatory variables. A linear relation between the (independently) variables and the (dependent) variable response is to model the multiple linear regression (MLR) purpose (Uyanık and Güler, 2013).

Multiple regression is essentially the expansion of the normal regression (OLS), which includes more than one explanatory variable.



~~~
```{r}
mlr <- lm(Output ~ Ease.and.convenient+Self.Cooking+ More.restaurant.choices +
Unaffordable+More.Offers.and.Discount +Good.Food.quality+Good.Tracking.system+Bad.past.experience+Age+
Maximum.wait.time, data = final_data)
summary(mlr)
```
~~~

```
Call:
lm(formula = Output ~ Ease.and.convenient + Self.Cooking + More.restaurant.choices +
    Unaffordable + More.Offers.and.Discount + Good.Food.quality +
    Good.Tracking.system + Bad.past.experience + Age + Maximum.wait.time,
    data = final_data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0997 -0.2168  0.1264  0.2348  0.6060

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.067735   0.192305   5.552 5.33e-08 ***
Ease.and.convenient      -0.021853   0.016217  -1.348   0.1786
Self.Cooking              0.080824   0.018819   4.295 2.23e-05 ***
More.restaurant.choices  -0.002572   0.014751  -0.174   0.8617
Unaffordable              0.035697   0.015017   2.377   0.0179 *
More.Offers.and.Discount -0.023428   0.015091  -1.552   0.1214
Good.Food.quality         0.015291   0.015416   0.992   0.3219
Good.Tracking.system     -0.016732   0.013996  -1.195   0.2326
Bad.past.experience       0.037585   0.017316   2.171   0.0306 *
Age                      -0.032941   0.006482  -5.082 5.89e-07 ***
Maximum.wait.time         0.109344   0.020035   5.458 8.76e-08 ***
```

From the results table above, we see the variables Self Cooking, Maximum Wait Time, Unaffordable, Age, and Bad Experience are statistically significant because the p-value is less than or equal to 0.05. Most of the features have positively impacted on the online food delivery experience, except Age. This shows that the older customers are, the less they use online food delivery services.

## 3.4 Principal Component Analysis

In addition to a logistical regression, the principal components analysis is the most common method of accelerating a learning algorithm (PCA). If the input dimension is too high, PCA can be a good alternative if you want it to be speeded up by reducing dimensions. Principal

Component Analysis, or PCA, is a dimensionality-reduction technique for reducing the dimensionality of large data sets by translating a large number of variables into a smaller one that retains the majority of the information in the large set (Abdi and Williams, 2010).

Naturally, reducing the number of variables in a data set reduces precision; however, the trick to dimensionality reduction is to substitute some accuracy for simplicity. Smaller data sets are simpler to experiment and imagine, and machine learning algorithms can analyze data even more easily and quickly without having to deal with extraneous variables. This is potentially PCA's most frequent use.

PCA is carried out by scale, so before applying PCA, we must assess the characteristics of our results. The scale function is used to standardize the features of a data set on unit size (mean = 0 and variance = 1), a prerequisite for many machine learning algorithms to optimize their output.

The original data has 10 independent variables, which are (1) Ease and convenience, (2) Self Cooking, (3) Maximum wait time, (4) More restaurant choices, (5) Unaffordable, (6) More Offers and Discount, (7) Good Food quality, (8) Good Tracking system, (9) Age, (10) Bad past experience). In this section, the function princomp() will be used to carry out a principal component analysis.

```
scaled_data <- scale(final_data)
fit <- princomp(scaled_data, cor = TRUE)
summary(fit)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4    Comp.5
## Standard deviation     1.3831715 1.3000006 1.1076103 1.01827986 0.9845411
## Proportion of Variance 0.1739239 0.1536365 0.1115273 0.09426308 0.0881201
## Cumulative Proportion  0.1739239 0.3275604 0.4390878 0.53335086 0.6214710
##                            Comp.6     Comp.7     Comp.8     Comp.9   Comp.10
## Standard deviation     0.94600967 0.87938899 0.85898311 0.81114123 0.7496328
## Proportion of Variance 0.08135766 0.07030227 0.06707745 0.05981365 0.0510863
## Cumulative Proportion  0.70282862 0.77313089 0.84020835 0.90002199 0.9511083
##                           Comp.11
## Standard deviation     0.73335447
## Proportion of Variance 0.04889171
## Cumulative Proportion  1.00000000
```

It is clear that the first 6 principal components account for about 70% of total data variance. Thus, the loadings of the first 6 principal components are collected as below.

```
##                             Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## Ease.and.convenient           0.29   0.34   0.24   0.16   0.30   0.13
## Self.Cooking                  0.38  -0.10   0.20   0.51   0.25  -0.18
## More.restaurant.choices       0.29   0.40   0.09  -0.10   0.19  -0.21
## Unaffordable                  0.36  -0.31  -0.39   0.21   0.05   0.29
## More.Offers.and.Discount      0.27   0.30  -0.43   0.23  -0.11  -0.36
## Good.Food.quality             0.29   0.20   0.17  -0.40   0.24   0.57
## Good.Tracking.system          0.24   0.21  -0.42  -0.50  -0.04  -0.24
## Bad.past.experience           0.39  -0.20  -0.21   0.01  -0.43   0.34
## Age                          -0.29   0.02  -0.48   0.22   0.55   0.24
## Maximum.wait.time             0.05  -0.45  -0.12  -0.36   0.50  -0.28
## Output                        0.33  -0.45   0.26  -0.11   0.04  -0.22
```

# 4 Conclusion

The paper examines the relationship between the online food ordering decision of people in Bangalore with various factors of orderers, such as age, the preferences of self-cooking, ease and convenience of online food, food quality, tracking system, or past experiences. Using some multivariate techniques (consists of multivariate normality test, logistic regression, multiple linear regression, and principal component analysis), the paper finds a significant relationship between online food purchasing decision with age, preferences of self-cooking, waiting time, level of affordability, and bad past experience. The implication of this result may help participants in the online food industry have a better understanding and insights into the preference of online food customers, in order to have an appropriate strategy to gain and retain more customers for this industry. However, there should be more variables and considerations added to the analysis, such as including more relevant variables like delivery, freshness, temperature, or missing items.

# 5 References

Ben Roshan. Online Food Delivery Preferences – Bangalore region. *Assessing online food delivery demand in Bangalore, India,* <https://www.kaggle.com/benroshan/online-food-delivery-preferencesbangalore-region>

Abdi, H. and Williams, L.J., 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), pp.433-459.

Frost and Sullivan 2019, *$9.6 Billion in Investments Spurring Aggressive Expansion of Food Delivery Companies,* viewed 12 April 2021, <https://ww2.frost.com/news/press-releases/9-6-billion-in-investments-spurring-aggressive-expansion-of-food-delivery-companies>

Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M. and Klein, M., 2002. *Logistic regression*. New York: Springer-Verlag.

Székely, G.J. and Rizzo, M.L., 2005. A new test for multivariate normality. *Journal of Multivariate Analysis*, *93*(1), pp.58-80.

Statista 2021, Online Food Delivery, viewed 12 April 2021, <https://www.statista.com/outlook/dmo/eservices/online-food-delivery/india>

Uyanık, G.K. and Güler, N., 2013. A study on multiple linear regression analysis. *Procedia-Social and Behavioral Sciences*, *106*, pp.234-240.