BDP200 Final Assignment

Group01: Nevil Nalinkumar Patel & Ran Arino

Professor: Anita (Mahnaz) Malekzadeh

Data Preparation Phase

1. Description of the Dataset Attributes

Our Dataset: in-vehicle coupon recommendation Data Set

- **destination**: One from ['No Urgent Place', 'Home', 'Work'].
- **passanger**: Who are the passengers in the car?, One from ['Alone', 'Friend(s)', 'Kid(s)', 'Partner'].
- **weather**: One from ['Sunny', 'Rainy', 'Snowy'].
- **temperature**: One from [55, 80, 30].
- **time**: One from ['2PM', '10AM', '6PM', '7AM', '10PM'].
- **coupon**: One from ['Restaurant(<20)', 'Coffee House', 'Carry out & Take away', 'Bar', 'Restaurant(20-50)'].
- **expiration**: Coupon expiration, One from ['1d', '2h'].
- **gender**: One from ['Female', 'Male'].
- **age**: One from ['21', '46', '26', '31', '41', '50plus', '36', 'below21'].
- **maritalStatus**: One from ['Unmarried partner', 'Single', 'Married partner', 'Divorced', 'Widowed'].
- **has_children**: One from [1, 0].
- **education**: One from ['Some college - no degree', 'Bachelors degree', 'Associates degree', 'High School Graduate', 'Graduate degree (Masters or Doctorate)', 'Some High School']
- **occupation**: One from ['Unemployed', 'Architecture & Engineering', 'Student', 'Education&Training&Library', 'Healthcare Support', 'Healthcare Practitioners & Technical', 'Sales & Related', 'Management', 'Arts Design Entertainment Sports & Media', 'Computer & Mathematical', 'Life Physical Social Science', 'Personal Care & Service', 'Community & Social Services', 'Office & Administrative Support', 'Construction & Extraction', 'Legal', 'Retired', 'Installation Maintenance & Repair', 'Transportation & Material Moving', 'Business & Financial', 'Protective Service', 'Food Preparation & Serving Related', 'Production Occupations', 'Building & Grounds Cleaning & Maintenance', 'Farming Fishing & Forestry']
- **income**: One from ['$37500 - $49999', '$62500 - $74999', '$12500 - $24999', '$75000 - $87499', '$50000 - $62499', '$25000 - $37499', '$100000 or More', '$87500 - $99999', 'Less than $12500']
- **car**: One from [nan, 'Scooter and motorcycle', 'crossover', 'Mazda5', 'do not drive', 'Car that is too old to install Onstar :D']
- **Bar**: How many times do you go to a bar every month?, One from ['never', 'less1', '1~3', 'gt8', nan, '4~8']
- **CoffeeHouse**: How many times do you go to a coffeehouse every month?, One from ['never', 'less1', '4~8', '1~3', 'gt8', nan]
- **CarryAway**: How many times do you get take-away food every month?, One from [nan, '4~8', '1~3', 'gt8', 'less1', 'never']

- **RestaurantLessThan20**: how many times do you go to a restaurant with an average expense per person of less than $20 every month?, One from ['4~8', '1~3', 'less1', 'gt8', nan, 'never']
- **Restaurant20To50**: how many times do you go to a restaurant with average expense per person of $20 - $50 every month?, One from ['1~3', 'less1', 'never', 'gt8', '4~8', nan].
- **toCoupon_GEQ5min**: Is driving distance to the restaurant/bar for using the coupon greater than 5 minutes?, every participant answered [1].
- **toCoupon_GEQ15min**: Is driving distance to the restaurant/bar for using the coupon greater than 15 minutes?, One from [0, 1].
- **toCoupon_GEQ25min**: Is driving distance to the restaurant/bar for using the coupon greater than 25 minutes?, One from [1].
- **direction_same**: Is the restaurant/bar in the same direction as your current destination?, One from [0, 1].
- **direction_opp**: Is the restaurant/bar in the same direction as your current destination?, One from [1, 0].
- **Y**: Is the coupon accepted?, One from [0, 1].

In this dataset, the target(dependent) variable is **Y**, and other attributes are description(independent) variables. In addition, all independent variables are categorical variables.

2. Description of Each Code

(1) Rename Some Variable Names

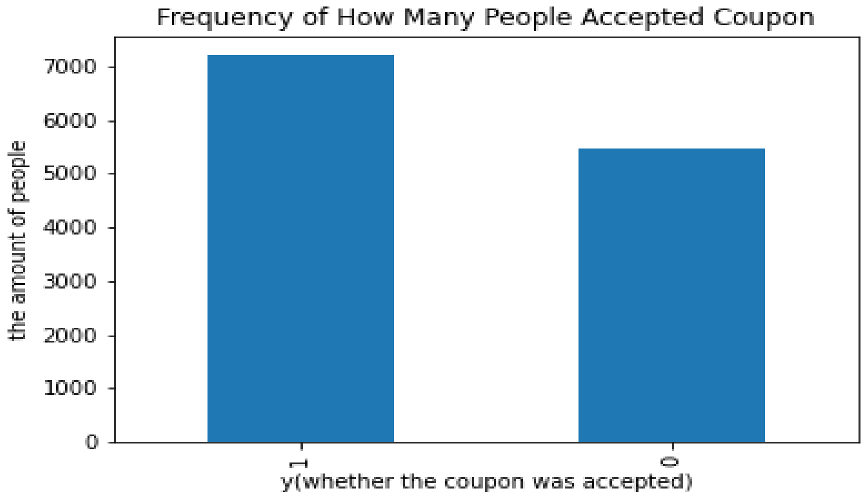
As we coded in the Jupyter Notebook file, we changed some original variables into a unified name with a small letter and underline and as short as possible. The following are the changed variable names.

> ['destination', 'passenger', 'weather', 'temperature', 'time', 'coupon', 'expiration', 'gender', 'age', 'marry', 'children', 'education', 'occupation', 'income', 'car', 'bar', 'cafe', 'takeout', 'restaurant_less20', 'restaurant_less50', 'coupon_dist_5', 'coupon_dist_15', 'coupon_dist_25', 'direction_same', 'direction_opp', 'y']

(2) Univariate & Bivariate Analysis

In this phase, we will observe the frequency of each variable (univariate analysis), which includes independent and dependent variable(s). Simultaneously, we will check how the individual value of each independent variable affects the dependent variable by plotting a stacked bar chart (bivariate analysis). In other words, we will conduct the univariate analysis of each dependent variable and its bivariate analysis with an independent variable at the same time in order to specify which dependent variables would be critical to predict an independent variable.

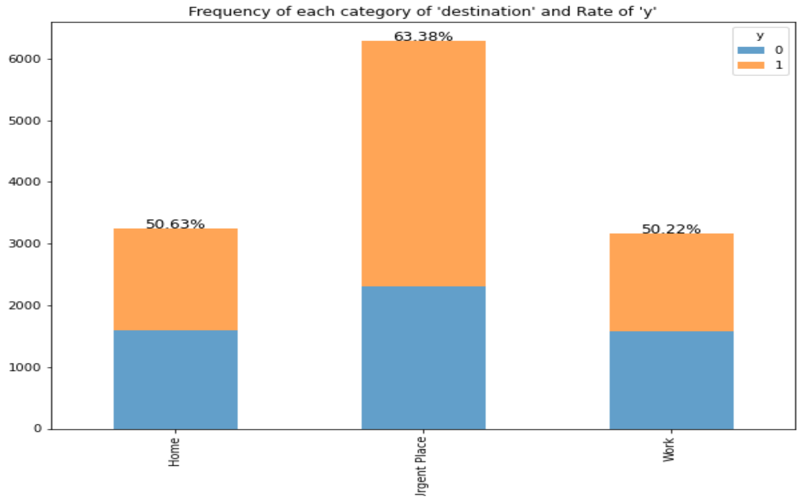
(2)-1: Frequency of Y (how many people accepted coupon)



This graph shows that a dependent (target) variable is distributed well-balanced without indicating the extreme majority or minority group(s). Therefore, we would expect that the trained machine is less like to overreact to a certain value (usually this is the majority value in a variable).

In addition, as we showed in the Jupyter Notebook, the rate of “y=1”(accept) and “y=0”(non-accept) is 56.84% and 43.16, respectively. We will analyze the bivariate analysis by utilizing this bise ratio.

(2)-2: The relationship between “destination” & “y”



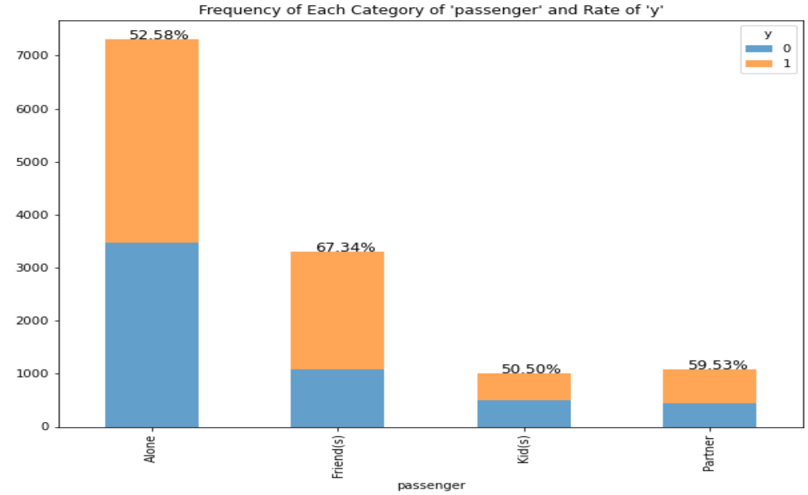
For Univariate Analysis:

- The number of people who answered “No Urgent Place” is the most frequent category.
- We find its frequency of each category a little bit imbalanced, but it’s not extreme one.

For Bivariate Analysis:

- The “Home” and “Work” categories in a ‘destination’ attribute may decrease the case where people accept coupon because the ratio of “y=1” changes from 56.84% (base) to around 50.00%.
- On the other hand, for those who answered ‘Not Urgent Place’, the probability to accept the coupon increased modestly compared to the base ratio and those who answered the other two types – 63% of the people who answered ‘Not Urgent Place’ accepted the coupon.
- Therefore, this independent variable could affect the outcome of the target variable.

(2)-3: The relationship between “passenger” & “y”



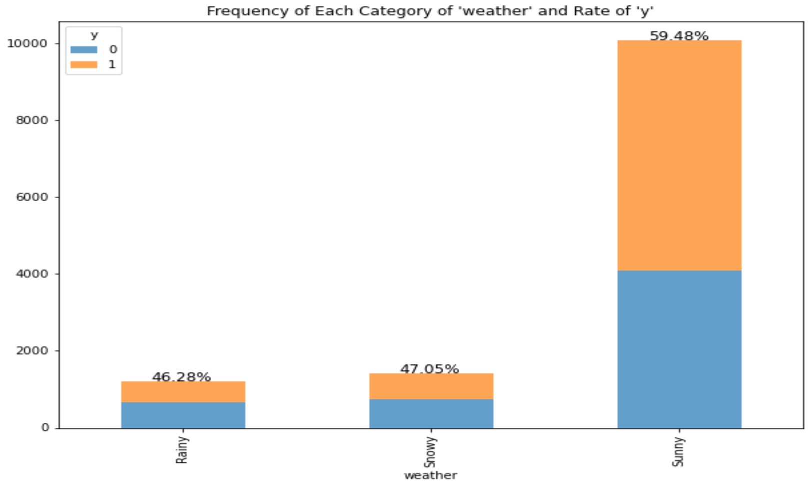
For Univariate Analysis:

- This shows the passengers in the car when participants were asked.
- The “Alone” is the most frequent, and the overall frequency is skewed towards its category.

For Bivariate Analysis:

- When the passengers were friends, the probability of accepting coupons is higher than others and also the base ratio of “y=1”.
- Additionally, the difference between the highest rate of “y” (67.34%) and the lowest rate of “y” (50.50), each category of this independent variable could affect the target variable, especially those who are with friends in the car are likely to accept coupons in the future.

(2)-4: The relationship between “weather” & “y”



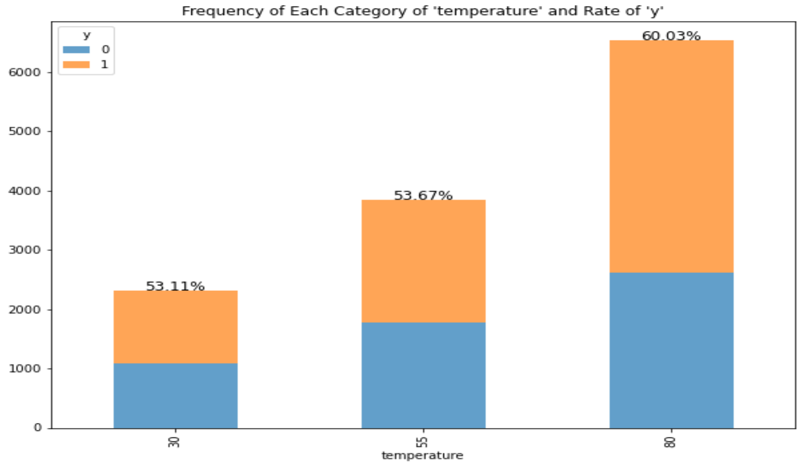
For Univariate Analysis:

- This shows the Frequency of the earning during the different weather.
- The sunny weather has the highest earning for vehicle coupons.
- However, the distribution is extremely skewed towards sunny days, which shows that most researching dates were good weather.

For Bivariate Analysis:

- During rainy weather, the rate of y is less than 50% and same goes for the snowy weather, so we can conclude that both weather days could decrease the coupon acceptability (because the base ratio of “y=1” is around 56%).
- If we talk about the rate of y for sunny then it is at least 10% higher than both rainy and snowy
- The rate of y for sunny weather is the highest rate of y (59.48%) and the lowest rate of y (46.28%) in each category of this independent variable could affect the target variable.

(2)-5: The relationship between "temperature" & "y"



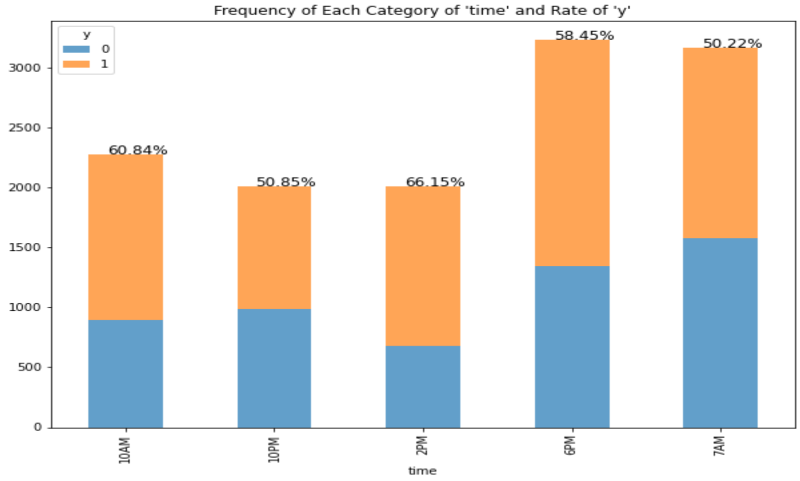
For Univariate Analysis:

- First of all, this shows the temperature when the questionnaire was conducted, but it doubts why this independent variable is categorical one.
- If this independent variable includes the training data for machine learning, the future analysis or prediction could cause bias because only three temperatures would be trained by the machine.
- Therefore, this independent variable should be removed.

For Bivariate Analysis:

- If the temperature increases, the rate of accepting coupons may increase as well.
- However, there are not enough categories (kinds of temperature), so we cannot conduct the appropriate data analysis.

(2)-6: The relationship between "time" & "y"



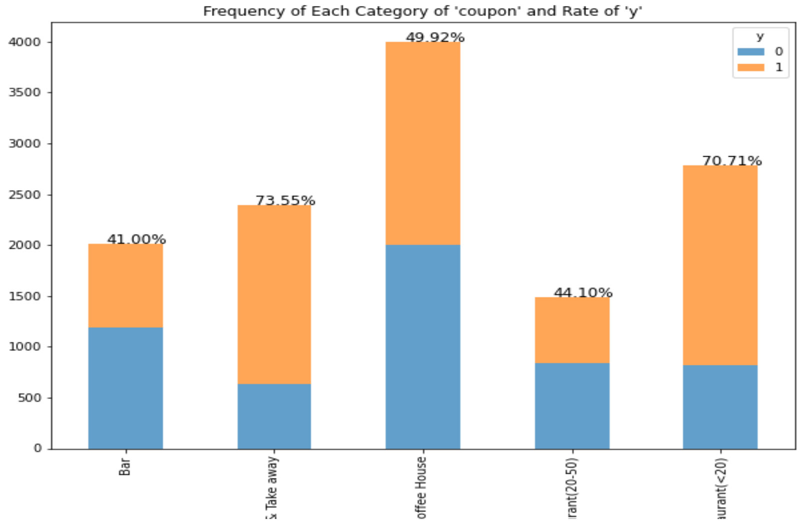
For Univariate Analysis:

- The balance of each category is not bad.
- It means that each category is not excessively skewed towards a certain category.

For Bivariate Analysis:

- This bivariate analysis indicates that the ratio of “y=1” increased in two categories (10 AM and 2 PM), its ratio was unchanged at 2 PM, and at 7 AM and 10 PM decreased its ratio, compared to the base ratio of “y=1”, which is 56.84%.
- This independent variable could have a good predictability to determine whether people accept coupons or not.

(2)-7: The relationship between “coupon” & “y”



For Univariate Analysis:

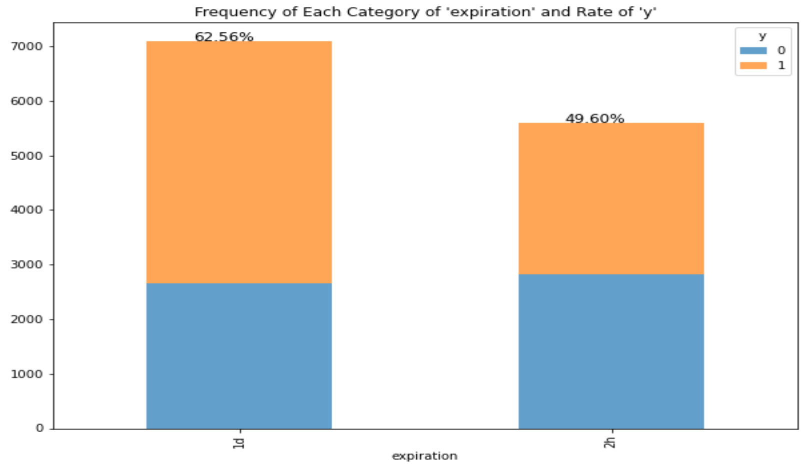
- This graph shows what kinds of coupons were provided to participants during the survey.
- The structure is mostly dependent on coffee house, so it means that the coupon for coffee shop were provided frequently.
- People would also be interested in what kinds of coupons are most acceptable for people, so researchers should have controlled the population of each coupon, which means that each coupon was provided to participants equally such as unifying as 1000 coupons and 5000 in total.

For Bivariate Analysis:

- This bivariate analysis indicates that the ratio of “y=1” increased in two categories (take away and restaurant (<20), more specifically, both increased by around 15% (base ratio was around 56%).
- The coupons for Coffee House slightly declined the rate of coupons acceptability.
- On the other hand, if the coupons were the other two types – Bar and Restaurant (20-50), people are less likely to accept coupons because the rate of “y=1” declined from its base ratio (around 56%).

- This independent (kinds of coupons) variable would affect whether people accept coupons or not.

(2)-8: The relationship between “expiration” & “y”



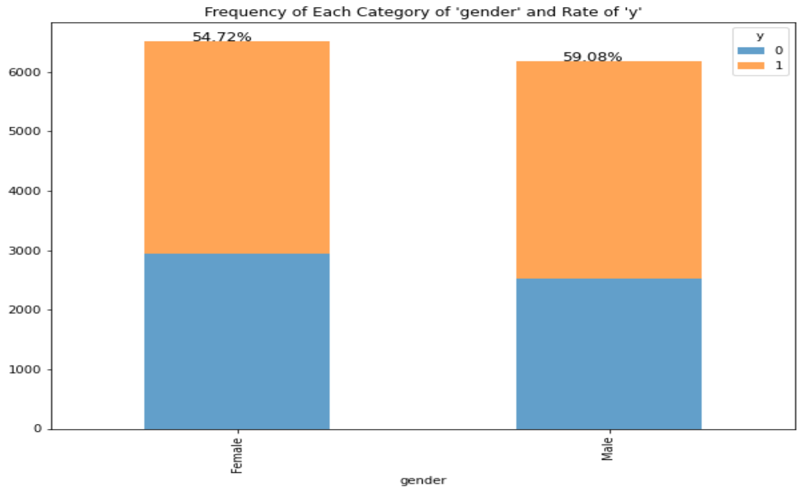
For Univariate Analysis:

- Apparently, each coupon had an expiration date, 1 day or 2 hours.
- As shown above, this distribution is not skewed toward one side: this graph is well balanced
- However, there may be shown more categories in the future.

For Bivariate Analysis:

- If we focus on the rate of accepting coupons in terms of each category 1d and 2h, we can confirm that there are slight differences between both.
- The ratio of y = 1 is slightly higher in category 1d.
- Therefore, this independent variable possibly affects the people’s determination of getting coupons or not.

(2)-9: The relationship between “gender” & “y”



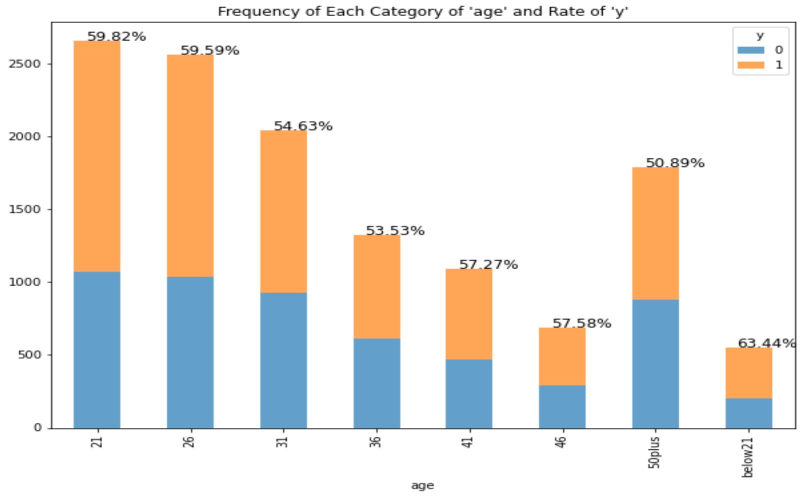
For Univariate Analysis:

- Well-balanced bar graph.

For Bivariate Analysis:

- If we focus on the rate of accepting coupons in terms of each category (female and male), we can confirm that there are no significant differences between both.
- Therefore, we can remove the independent variables.

(2)-10: The relationship between “age” & “y”



For Univariate Analysis:

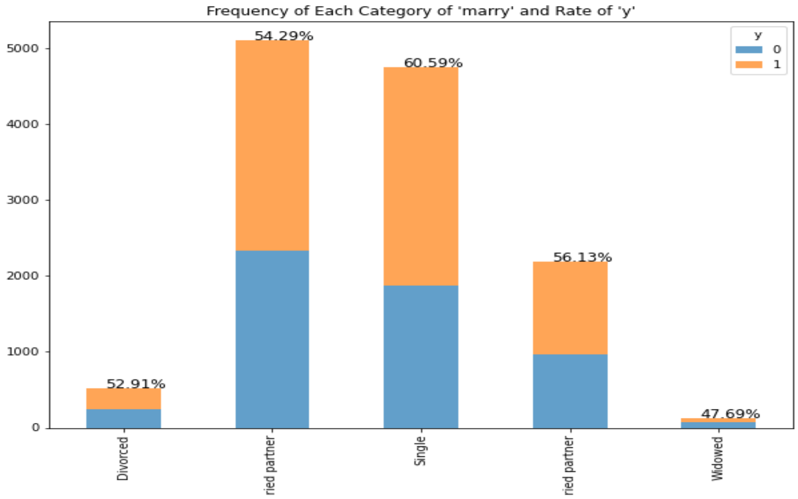
- Similar to the “temperature” variable, why are there so few different types of age?
- In the future questionnaire, we expect that a lot of different ages are recorded.

- We are interested in the impact of participants' age on the target variable, but it we make machine or computer train this independent variable, the performance of the machine prediction could deteriorate even more.
- Therefore, this independent variable will be removed from our training data.

For Bivariate Analysis:

- We evaluate that the rate of accepting coupons is constant at different ages: there is no significant difference among ages.

(2)-11: The relationship between "marry" & "y"



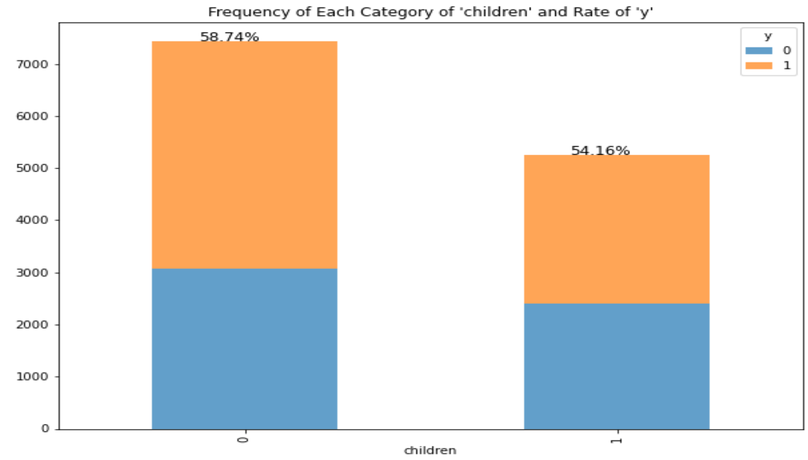
For Univariate Analysis:

- This indicates the participants' marital status.
- The category, "married partner" is the most frequent one, which is close to 5000 and a little bit higher than Single.
- The category, "Unmarried partner" is close to half of the amount of the "Married partner", and the other two categories are extremely low.

For Bivariate Analysis:

- There is no significant difference between marital status by focusing on the larger population of categories: married partner, single, and unmarried partner (we ignore the other two categories due to much less population).
- We may remove this independent variable.

(2)-12: The relationship between “children” & “y”



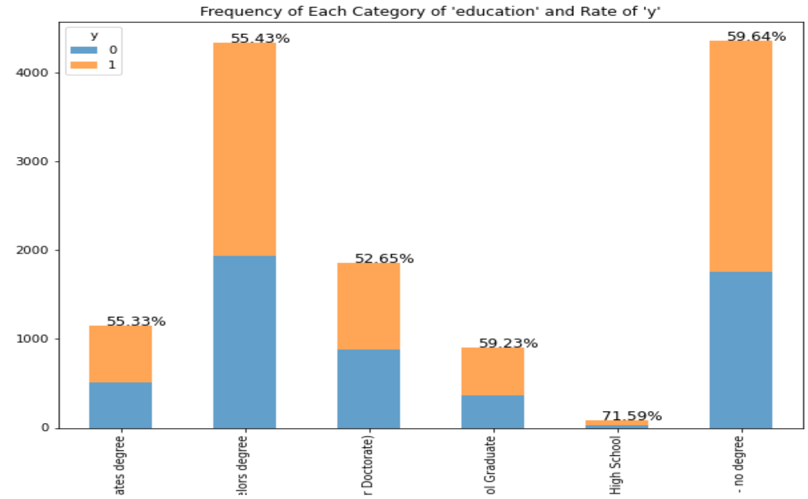
For Univariate Analysis:

- We can say that each category is well balanced.

For Bivariate Analysis:

- If we focus on the rate of accepting coupons in terms of each category of children (“0” and “1”), we can confirm that there are no significant differences between those two categories.
- Therefore, we can remove the independent variables.

(2)-13: The relationship between “education” & “y”



For Univariate Analysis:

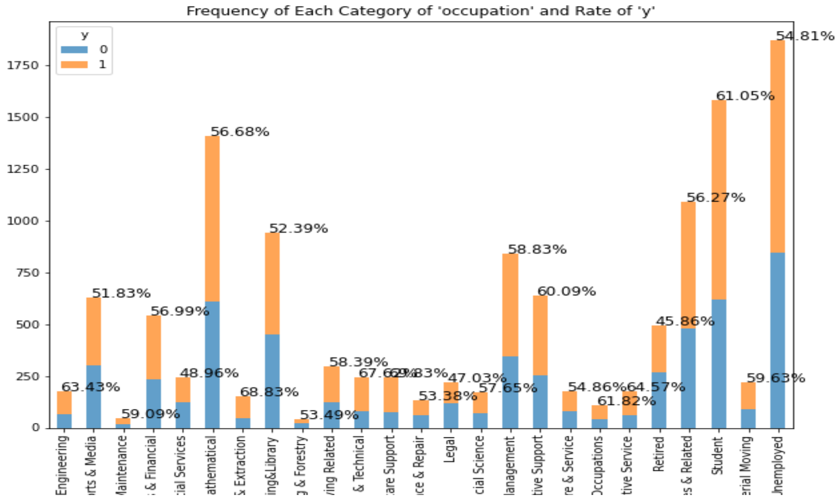
- This indicates the participants’ education.
- The category, “Bachelor's degree” is the most frequent one, which is close to 5000 and almost equal to no degree.

- The category, “High School” is close to zero, and the other three categories are less 2000.
- Therefore, each category of this independent variable is not well-balanced.

For Bivariate Analysis:

- Our concern is that the category ‘High School’ which was the highest rate of ‘y=1’ and the lowest population might cause basis.
- More specifically, if we make the machine learn the training dataset with this independent variable, its machine might learn that “people who answer High School in this section highly accept the coupons” mistakenly: the low population is less likely to indicate the fact.
- Therefore, we can remove this independent variable.

(2)-14: The relationship between “occupation” & “y”



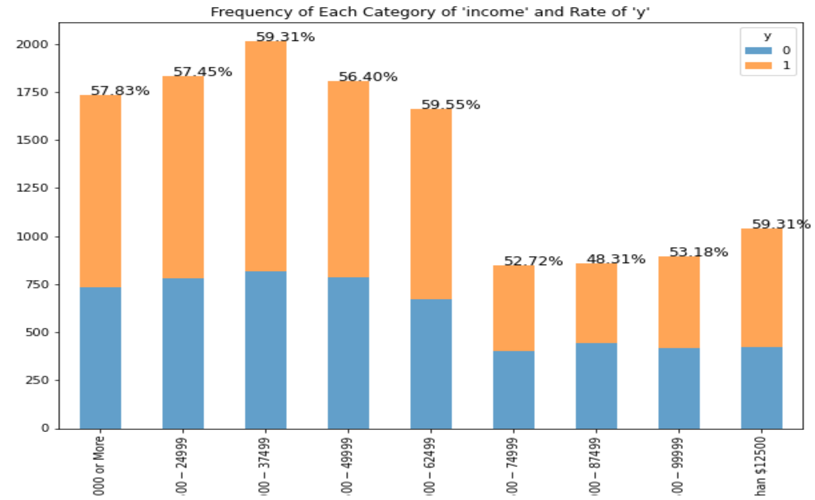
For Univariate Analysis:

- Obviously, there are too many categories.
- We will evaluate the univariate and bivariate analysis again after rearranging its categories: Unemployed (including Retired), Students, and Worker.

For Bivariate Analysis:

- Re-evaluate later.

(2)-15: The relationship between “income” & “y”



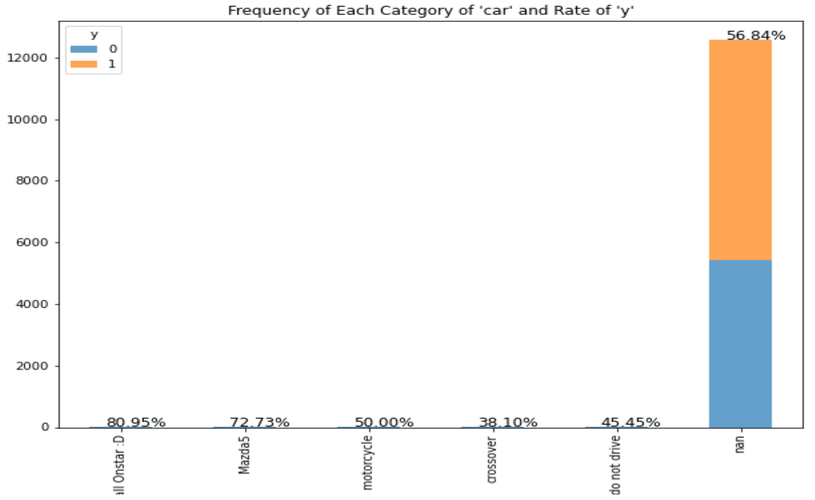
For Univariate Analysis:

- The category concentrates on the relatively lower income (12500-62499).
- On the other hand, the number of people whose income is $100000 or more is increased by twice compared to the previous income layers (between $62500 and $99999).
- One of the possibilities, some outliers are included in the “$100000 or more” category.
- However, this is not a numerical variable, so we cannot deal with outliers appropriately.

For Bivariate Analysis:

- From this graph, we cannot investigate that the higher or lower the individuals’ income, the higher the probability that they accept the coupons.
- We are interested in the relationship between the income and the acceptance of coupons, we might remove this independent variable from the training set.

(2)-16: The relationship between “car” & “y”



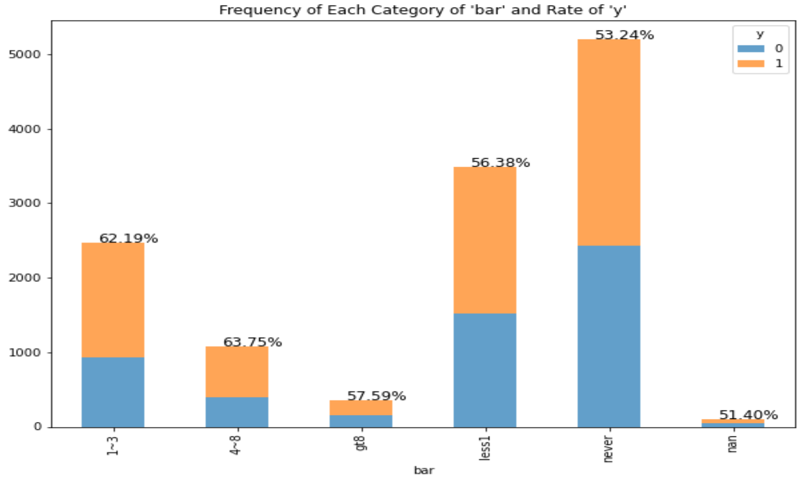
For Univariate Analysis:

- There are a lot of missing values, so we must remove this independent variable.

For Bivariate Analysis:

- No mention

(2)-17: The relationship between “bar” & “y”



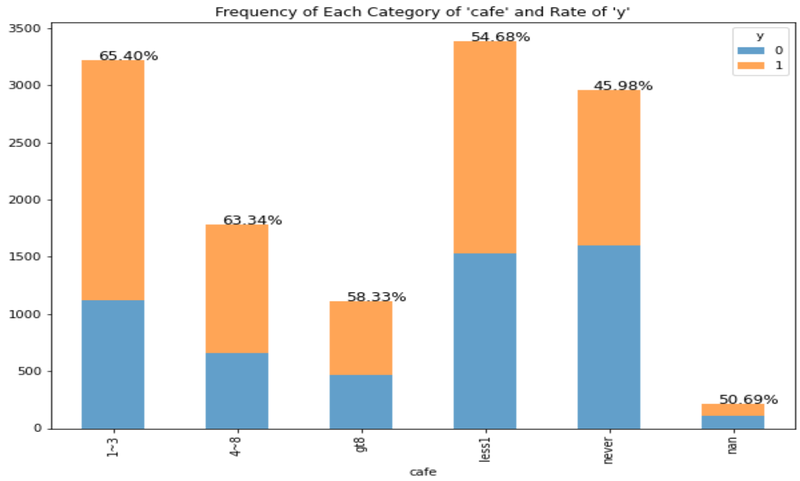
For Univariate Analysis:

- This shows how many times each participant goes to a bar every month.
- The missing values are few, so we can replace it on the most frequent category, which is “never”.
- Also, considering the few numbers of people in a “gt8” category, we can aggregate it and “4-8” into “gt4”.

For Bivariate Analysis:

- After arranging some categories and replacing the missing value, we will investigate the bivariate analysis.

(2)-18: The relationship between “cafe” & “y”



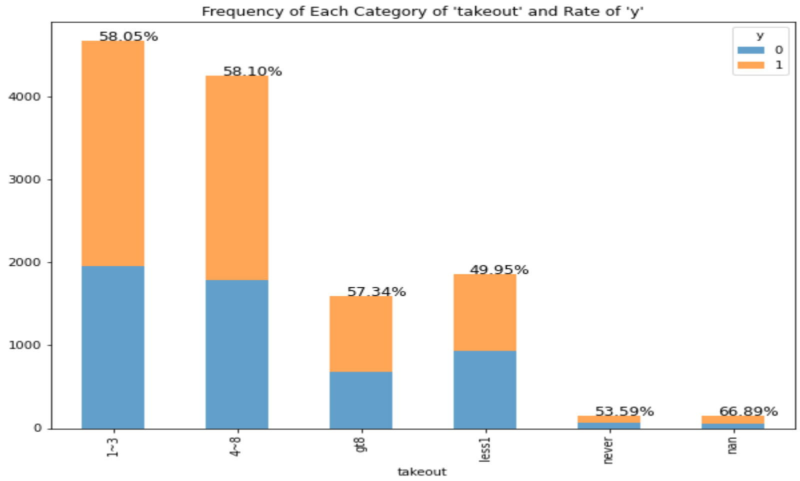
For Univariate Analysis:

- This shows how many times each participant goes to the Coffee House every month.
- In terms of the missing values and a category of “gt8”, we can conduct the same strategies as the previous independent variable (“Bar”).

For Bivariate Analysis:

- Assessing it after organizing the categories.

(2)-19: The relationship between “takeout” & “y”



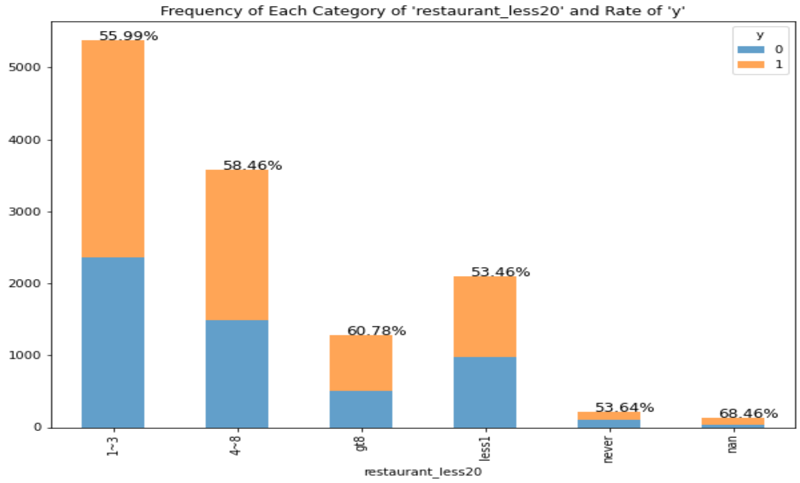
For Univariate Analysis:

- This shows how many times each participant gets take-away food every month.
- In terms of the missing values and a category of “gt8”, we can conduct the same strategies as the previous independent variables (“Bar”, “Cafe”)

For Bivariate Analysis:

- Assessing it after organizing the categories.

(2)-20: The relationship between “restaurant_less20” & “y”



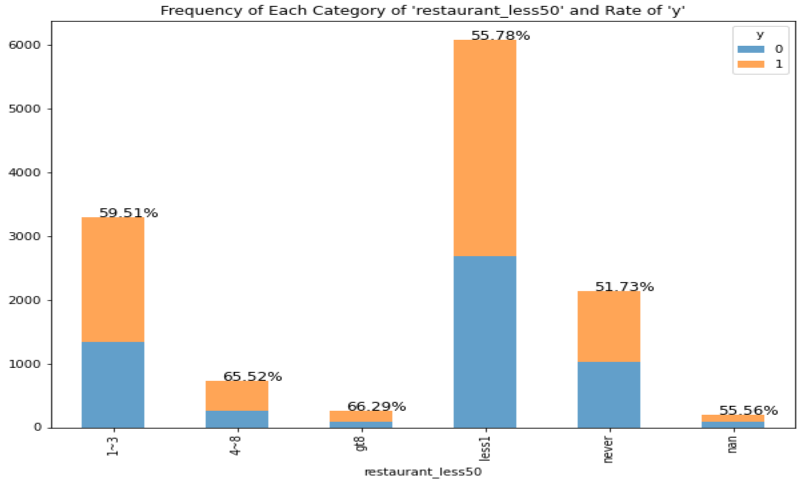
For Univariate Analysis:

- This shows how many times each participant goes to a restaurant with an average expense per person of less than $20 every month.
- In terms of the missing values and a category of “gt8”, we can conduct the same strategies as the previous three independent variables.

For Bivariate Analysis:

- Assessing it after organizing the categories.
-

(2)-21: The relationship between “restaurant_less50” & “y”



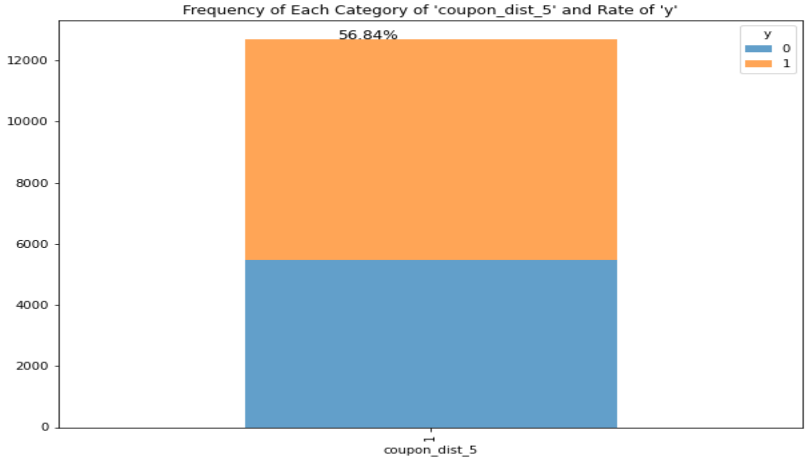
For Univariate Analysis:

- This shows how many times each participant goes to a restaurant with an average expense per person of less than $50 every month.
- The categories in this independent variable tend to be skewed toward “less1” and “1-3”.
- In terms of the missing values and a category of “gt8”, we can conduct the same strategies as the previous three independent variables.

For Bivariate Analysis:

- Assessing it after organizing the categories.

(2)-22: The relationship between “coupon_dist_5” & “y”



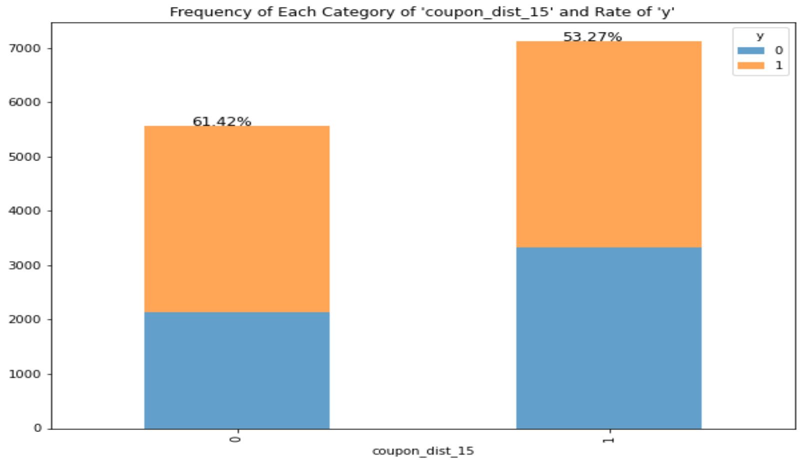
For Univariate Analysis:

- This outcome indicates that all coupons offered to participants were able to be used at locations more than 5 minutes away by car.
- We can remove this independent variable because it is a constant and doesn’t affect the dependent variable.

For Bivariate Analysis:

- None

(2)-23: The relationship between “coupon_dist_15” & “y”



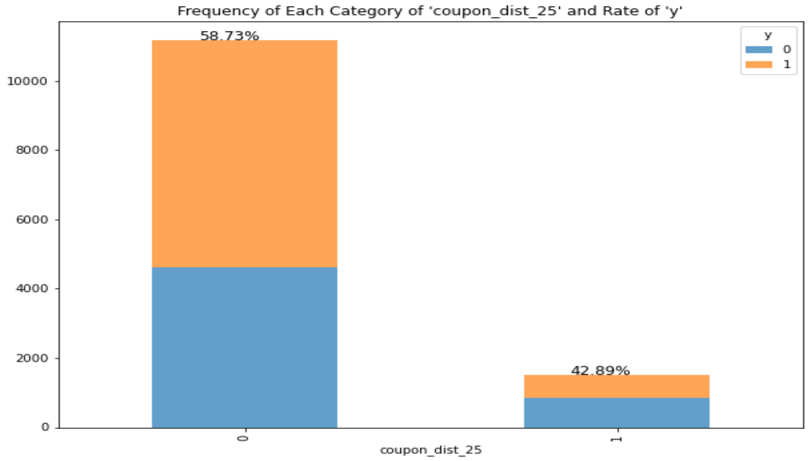
For Univariate Analysis:

- This independent variable shows whether the coupons were able to be used within 15 minutes away by car.
- In other words, “0” shows not greater than 15 min (= able to use it less than or equal to 15 min), and “1” shows greater than 15 min.
- The balance is not bad.

For Bivariate Analysis:

- As shown above, if people can use coupons at a place within less than or equal to 15 min away by car, they are likely to accept its coupons: its ratio increased by around 5% compared to its base ratio (around 56%).
- Also, when the distance is more than 15 min by car, people are less likely to accept coupons.
- Therefore, this independent variable could be a good dataset.

(2)-24: The relationship between “coupon_dist_25” & “y”



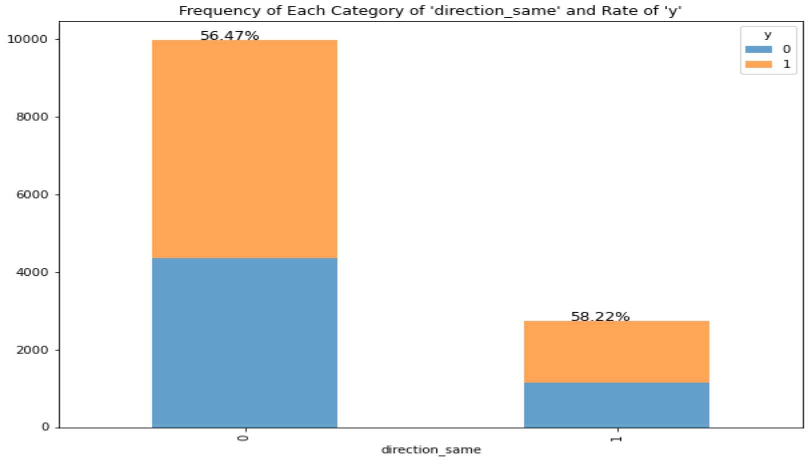
For Univariate Analysis:

- This independent variable shows whether the coupons were able to be used within 25 minutes away by car.
- In other words, “0” shows not greater than 25 min (= able to use it less than or equal to 25 min), and “1” shows greater than 25 min.
- Each category of this independent variable is not well-balanced.

For Bivariate Analysis:

- If the coupons can be used less than or equal to 25 min by car, the acceptability of coupons slightly increases.
- On the other hand, if not, its ratio clearly declined. (From around 56% initially to 42.89%)
- Therefore, this independent variable would be a great data to predict whether people will accept coupons or not in the future.

(2)-25: The relationship between “direction_same” & “y”



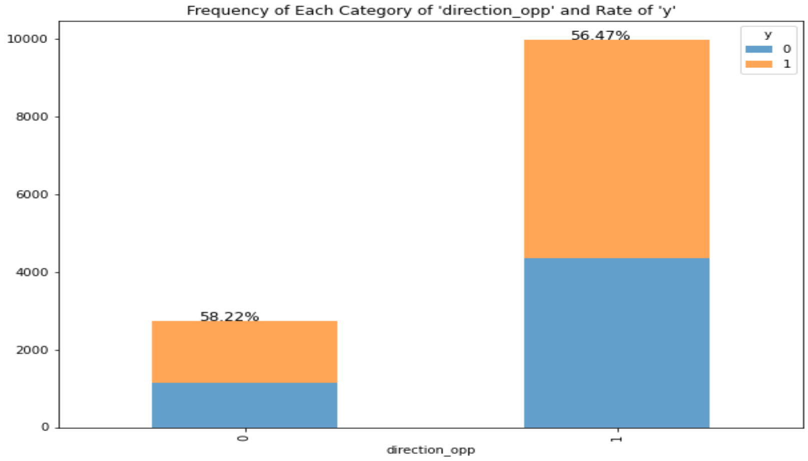
For Univariate Analysis:

- This independent variable shows whether the usage place of coupons is the same direction as each person’s current destination or not (“0”: opposite, “1”: same).
- Each category is not well-balanced, but it won’t control as long as researchers select the participants randomly.

For Bivariate Analysis:

- Considering each rate of “y=1”, this independent variable might not affect whether people are likely to accept the coupons or not: both ratios are almost the same as the base ratio (around 56%).
- Therefore, we can remove this independent variable.

(2)-26: The relationship between “direction_opp” & “y”



For Univariate Analysis:

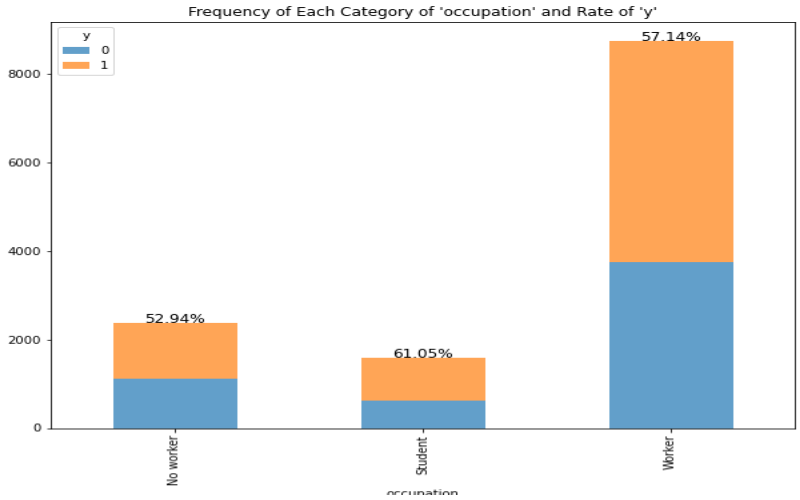
- We confirmed that this independent variable is exactly opposite to the previous independent variable, which is direction_same.
- Therefore, we can remove this independent variable due to the duplication.

For Bivariate Analysis:

- No mention

(3) Rearranged Category Name of Each Independent Variable & Apply Same Analysis

(3)-1: The relationship between "occupation" & "y" (rearranged)



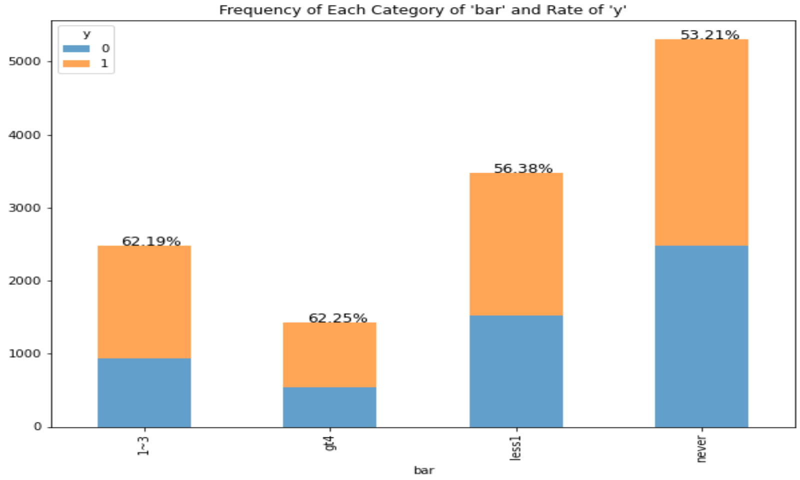
For Univariate Analysis:

- We aggregate most working occupations into "Worker", this category become the most frequent one.
- The graph indicates the skewness toward one category, but we think that aggregating many occupations is the appropriate method in order to pursue a simple prediction machine. For example, although more than 80% of a specific occupation such as teachers accepted coupons (the possibility to show a strong relationship), it may cause bias if we have few samples, such as only 10 teachers among 10000 participants.
- In other words, we should doubt that few samples are likely to indicate the fact in the real world.

For Bivariate Analysis:

- If participants were students, the rate of accepting coupons increases by around 5% compared to the dependent variable's basis (56.84).
- Although we cannot find significant effect for "Worker" and "No worker" categories, we may include this independent variable into the training dataset.

(3)-2: The relationship between “bar” & “y” (rearranged)



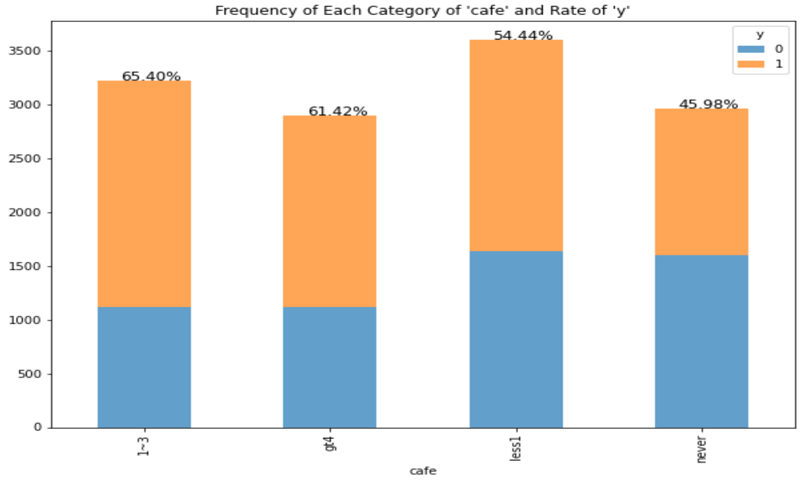
For Univariate Analysis:

- It's concerned that a category “gt4” is still few populations.

For Bivariate Analysis:

- As people can easily imagine, the ratio of accepting coupons tended to be higher among those who go to the bar every month than those who do not.
- More specifically, the categories of “1-3” and “gt4” are slightly higher than the base ratio (around 56%) of “y=1”, which indicates that people who go to bar more than 1 time per month are likely to accept coupons compared to those who not.

(3)-3: The relationship between “cafe” & “y” (rearranged)



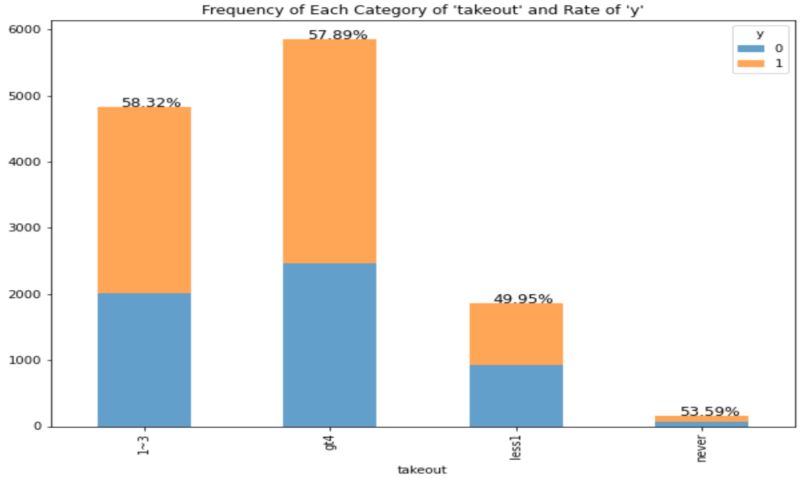
For Univariate Analysis:

- This bar graph shows that each category is well distributed.

For Bivariate Analysis:

- It difficult to understand why the ratio of “y=1” in a category of “1-3” is higher than the “gt4” category, but it’s certain that both are likely to encourage people accept coupons.
- Also, those who never go to the café in a month are less likely to acquire coupons.

(3)-4: The relationship between “takeout” & “y” (rearranged)



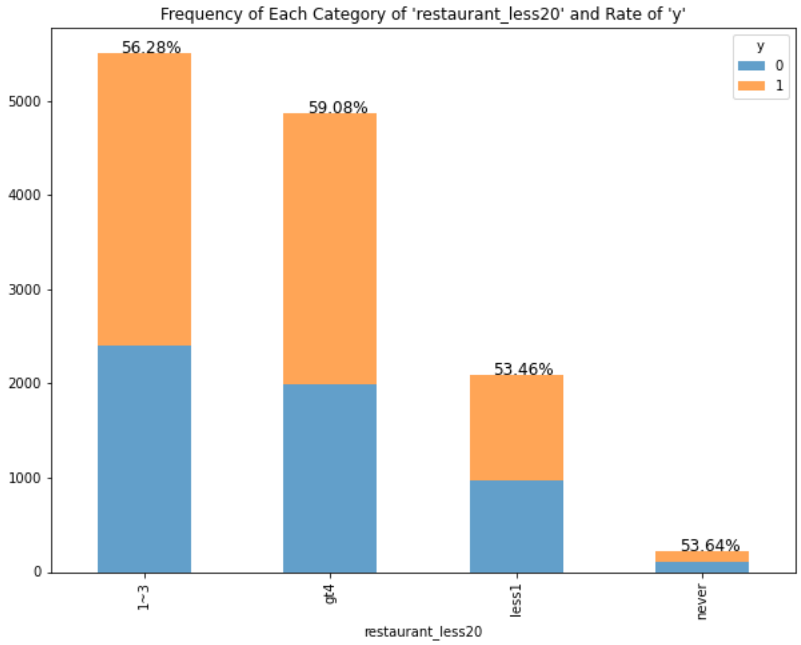
For Univariate Analysis:

- It's concerned that a category “never” is still few populations.

For Bivariate Analysis:

- The focus point is that those who answered “less1” is lower than the base ratio of accepting coupons (around 56%).
- Although those who use takeout service frequently did not show the significant effect compared to the base ratio of “y=1”, This independent variable may contribute to predicting the target variable in the future.

(3)-5: The relationship between “restaurant_less20” & “y” (rearranged)



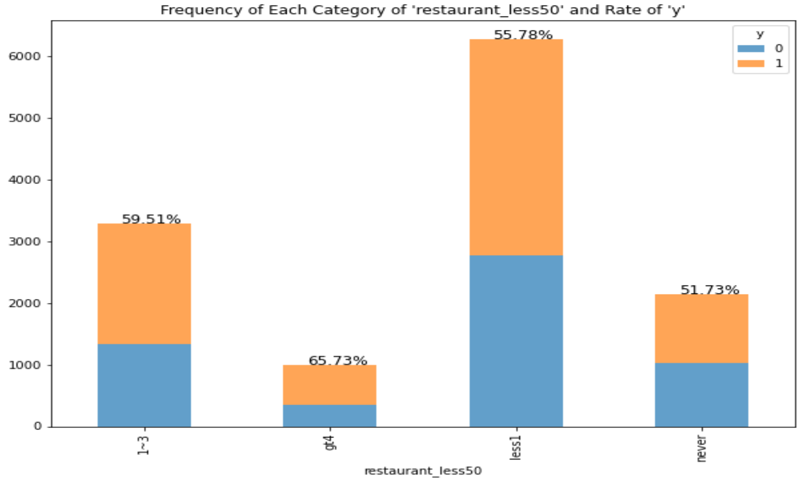
For Univariate Analysis:

- It's concerned that a category “never” is still few populations.

For Bivariate Analysis:

- People would guess that the more people go to reasonable restaurants, the higher the possibility that they accept getting coupons, but this graph doesn’t seem to prove it significantly.
- It might be controversial whether this independent variable includes the training dataset or not, but it’s worth analyzing and predicting the target variable based on including it the training set.

(3)-6: The relationship between “restaurant_less50” & “y” (rearranged)



For Univariate Analysis:

- It's concerned that the categories “gt4” and “never” are still few populations.

For Bivariate Analysis:

- Although there are few populations, the category “gt4” shows the highest ratio of accepting coupons of other categories.
- This result would broadly make sense from the human behavior: if people get discount coupons or special offers in expensive restaurants, they might want to go there because they cannot have lunch or dinner frequently (being relieved the barrier of the amount of money by coupons).

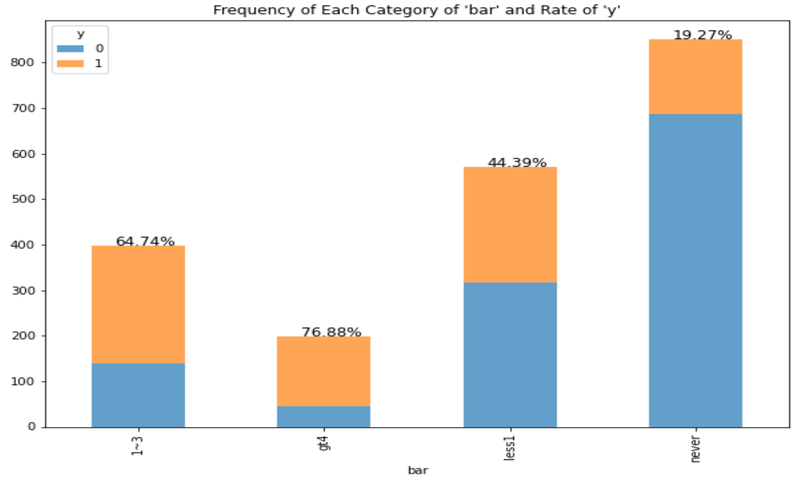
(4) Multivariate Analysis

In this phase, we will assess the percentage of accepting five kinds of coupons *["Bar", "Coffee House", Carry our & Take away", "Restaurant(<20)", and Restaurant(20-50)]* while focusing on how many times people go to each place in a month.

During those multivariate analyses, we would like to focus on the following two hypotheses (*A* in the below sentence indicates one place out of five ones):

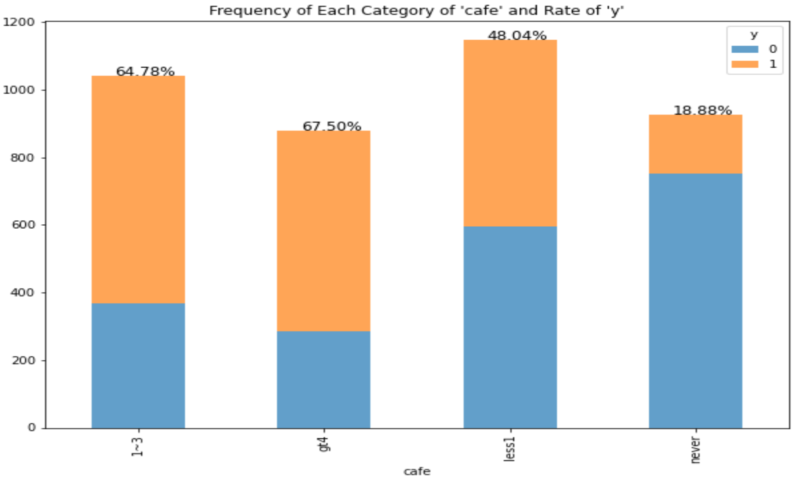
1. The more frequently people go to *A*, the higher the possibility that people accept *A*’s coupons.
2. Those who have never been to *A* are less likely to accept *A*’s coupons.

(4)-1: Ratio of accepting “Bar” coupons while focusing on how many times people go to the bar in a month.



- Considering the base ratio of accepting all kinds of coupons (56.84%) and the base rate of "Bar" coupon acceptability (41.00%), this outcome would strengthen our hypotheses.
- It is because the more frequently people go to bars in a month, the higher the probability that they will accept the bar coupons. Also, those who have never been to bars in a month are less likely to get the bar coupons.

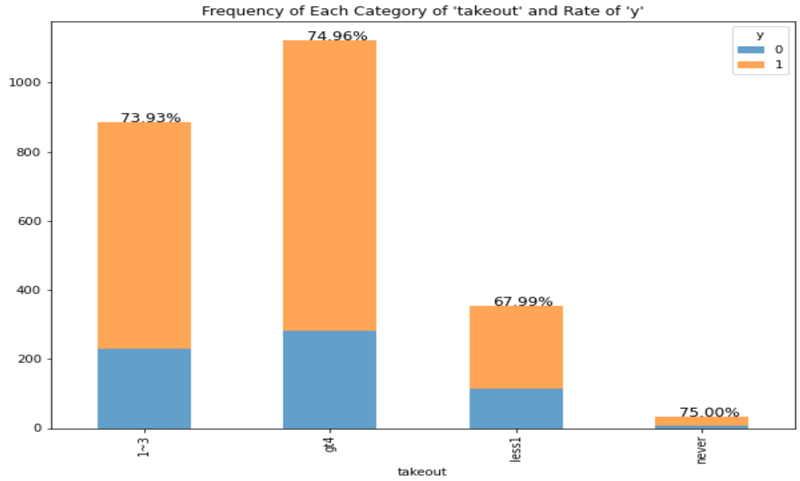
(4)-2: Ratio of accepting “Coffee House” coupons while focusing on how many times people go to the Coffee House (we changed its attribute name to “cafe”) in a month.



- Considering the base ratio of accepting all kinds of coupons (56.84%) and the base rate of "Coffee House" coupon acceptability (49.92%), this outcome would strengthen our hypotheses as well.

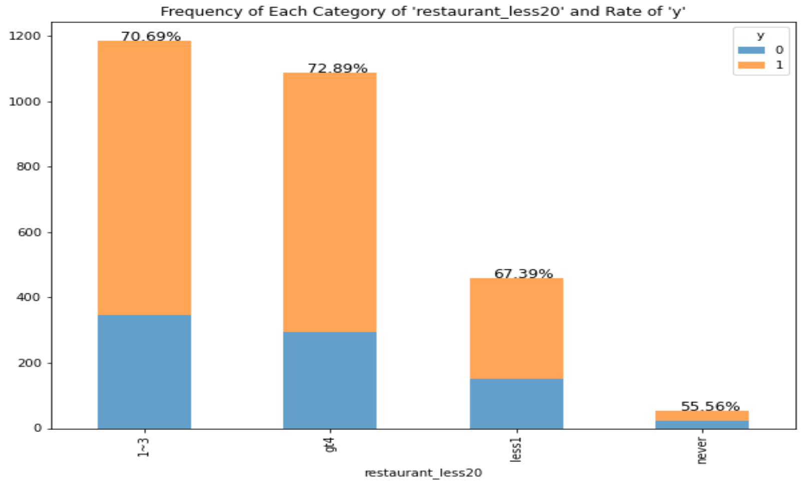
- It is because the more frequently people go to coffee houses (cafés) in a month, the higher the probability that they will accept the café coupons. Also, those who have never been to cafés in a month are less likely to get the café coupons.

(4)-3: Ratio of accepting “Carry out & Take away” coupons while focusing on how many times people go to the Carry out & Take away (we changed its attribute name to “takeout”) in a month.



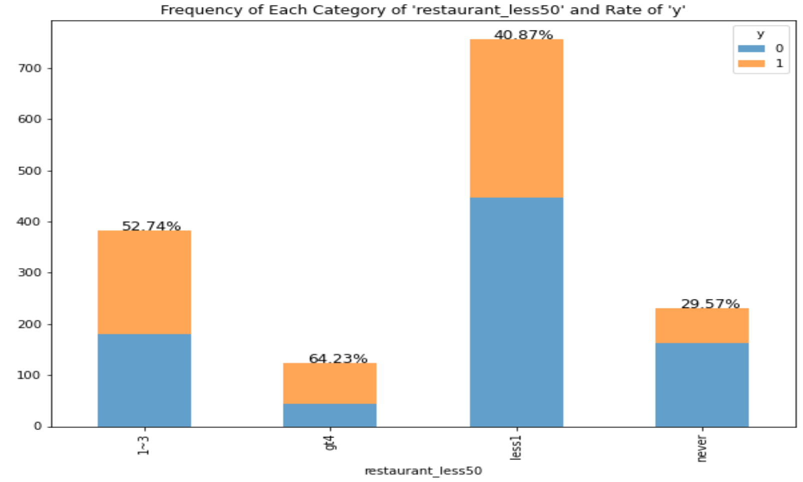
- Considering the base ratio of accepting all kinds of coupons (56.84%), this outcome would strengthen our hypotheses as well. However, by focusing on the base rate of "Carry out & Take away" (takeout) coupon acceptability (73.55%), we cannot say that this outcome provides us with a significant result.
- It is because if people frequently go to the takeout place (restaurant, café, etc.) more frequently in a month, their acceptability of its coupons is not significantly changing.
- Therefore, people might accept the takeout coupons regardless of their frequency of using the Carry out & Take away place.

(4)-4: Ratio of accepting “Restaurant(<20)” coupons while focusing on how many times people go to the Restaurant(<20) (we changed its attribute name to “restaurant_less20”) in a month.



- Similar to the previous outcome, considering the base ratio of accepting all kinds of coupons (56.84%), this outcome would strengthen our hypotheses as well. However, by focusing on the base rate of "Restaurant(<20) " (what we call “reasonable restaurant”) coupon acceptability (70.71%), we cannot say that this outcome provides us with a significant result.
- It is because if people go to the reasonable restaurant more frequently in a month, their acceptability of its coupons is not significantly changing: but those who have never been to reasonable restaurants are less likely to accept the coupons (corresponding it to our second hypothesis).
- Therefore, people might accept reasonable restaurant coupons regardless of their frequency of using those restaurants.

(4)-5: Ratio of accepting “Restaurant(20-50)” coupons while focusing on how many times people go to the Restaurant(20-50) (we changed its attribute name to “restaurant_less50”) in a month.



- Considering the base ratio of accepting all kinds of coupons (56.84%), this outcome would not support our first hypothesis but the second one is strengthened. It is because in terms of those who go to Restaurant(20-50) (what we call, expensive restaurants) a couple times in a month, their acceptability of coupons declined to 52.74% (around minus 4%).
- On the other hand, if we focus on the base ratio of accepting the coupons of expensive restaurants (44.10%), we can say that both our hypotheses are supported by this outcome.
- Those who have been to expensive restaurants more than a couple of times in a month are likely to accept coupons; whose ratios were higher than its base ratio (64.23% and 53.74% compared to 44.10%). Also, those who have never or been close to never been to expensive restaurants are less likely to accept the coupons of expensive restaurants.

In conclusion, it is likely that our hypotheses are partly proven, in particular, if the coupons are usable at “Bar”, “Coffee House”, and “Restaurant (20-50)”.

(5) Remove Some Independent Variables

Based on the univariate and bivariate analysis so far, we will remove the following variables:

> ['temperature', 'gender', 'age', 'marry', 'children', 'education', 'income', 'car', 'coupon_dist_5', 'direction_same', 'direction_opp']

And the remaining independent variables are shown in the following (“y” is a dependent variable):

> ['destination', 'passenger', 'weather', 'time', 'coupon', 'expiration', 'occupation', 'bar', 'cafe', 'takeout', 'restaurant_less20', 'restaurant_less50', 'coupon_dist_15', 'coupon_dist_25', 'y']

(6) Recheck the number of missing values

We have already replaced the missing values to the mode of each category, so we confirmed that there are no missing values. (Note: the reason why we used the mode to replace the missing value is that all of our independent variables are categorical variables.)

(7) Encoding each category of the independent variables

(7)-1: Checking the data type

As shown in the Jupyter Notebook, most independent variables are categorical variable with string type. In order to apply the Decision Tree method, we need to change them to the integer type.

Therefore, we will assign individual categorical variables to different numbers such 0, 1, 2 etc. (encoding).

(7)-2: Creating the function

In order to apply different numbers to each categorical variable(string) easily, we create the function. This function returns the dictionary: each key shows unique categories of each independent variable(string), and its value stands for the corresponding encoding number(integer).

(7)-3: Running the conversion

- Line 03: Starting the for loop. “c” will be assigned each column name one by one.
- Line 05: If each independent variable (column) is object type, Line 06 and Line 07 will be invoked.
- Line 06: Creating the encoding list by utilizing our define function: “create_encode”.
- Line 07: Running the conversion (encode)

As shown in the Jupyter Notebook, all variables in the dataset completed the conversion to numeric variables.

Experimental Design

In this phase, we split the data independent variables (shown as “X”) and an independent variable (shown as “y”).

After that we split the data (“X” and “y”) into the training and test data sets: the ratio of the training data is 70% (the test data is 30%).

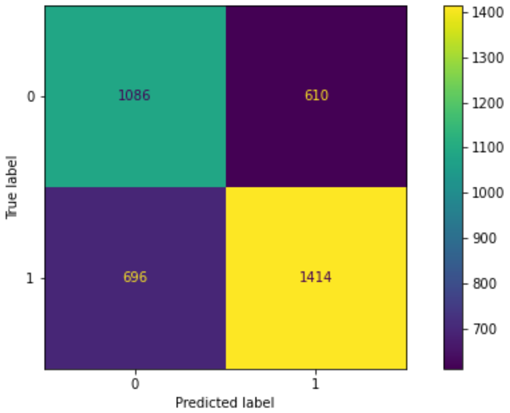
Modeling

In this phase, we established the machine learning model (we used Decision Tree Classification based on the “entropy” as its criterion).

After the computer learned the training data, we gave its trained machine to the new independent data (shown as X_test) to acquire the predicted y (shown as y_pred)

In the end, we plotted the confusion matrix by the test dependent variable (shown as y_test) and the predicted dependent variable (shown as y_pred). Below are the confusion matrix and the classification report provided by ‘matrics’ package in Python.

Confusion Matrix



Confusion Report

	Precision	Recall	F1-score	support
0	0.61	0.64	0.62	1696
1	0.70	0.67	0.68	2110
Accuracy	-	-	0.66	3806

Considering these outcomes, our machine learning model might increase the predictability of the target variable ‘y’ by only 10%. It is because the base ratio of 'y=1' was around 56%, which means that if we were to predict y (0 or 1) at random, we can guess whether a driver accepts coupons or not by 56% of accuracy ratio.

However, this case of accuracy (66%) is not necessarily a high degree of accuracy. Considering the Essential Steps of a Data Analytics Project that we learned, we should improve our machine learning model by changing the independent variables or adjusting the parameters for the next steps. However, these would be out of scope in this course. Although our data science project is obviously incomplete, we finalize our work since we fulfilled the criteria for the project of this BDP200 course.