**Berkeley – Machine Learning and Data Science**

**Overview:**

In this first practical application assignment of the program, you will seek to answer the question, “Will a customer accept the coupon?” The goal of this project is to use what you know about visualizations and probability distributions to distinguish between customers who accepted a driving coupon versus those who did not.

**Data:**

This data is from the UCI Machine Learning Repository and was collected via a survey on Amazon Mechanical Turk. The survey describes different driving scenarios, including the destination, current time, weather, and passenger, and then asks people whether they will accept the coupon if they are the driver. There are three possible answers people can choose from:

* “Right away”
* “Later, before the coupon expires”
* “No, I do not want the coupon”

The first two responses are labeled as “Y = 1,” and the third is labeled as “Y = 0.” There are five different types of coupons: Less expensive restaurants (under $20), coffee houses, carryout and takeaway, bars, and more expensive restaurants ($20–$50).

**Deliverables**

Your final submission should include the Practical Application 1 Jupyter Notebook and other relevant files, such as the dataset provided and uploaded to a public-facing GitHub repository as your first portfolio project. To explore the data, you will utilize your knowledge of pandas and Python to create statistical summaries demonstrating differences in those who accepted and rejected the coupon. Utilize the Matplotlib and Seaborn libraries to create a visualization using Python. Ensure that your findings are clearly stated in a separate section alongside actionable items and recommendations.

Your repository should also include a README document containing a brief nontechnical report that highlights the differences between customers who did and did not accept the coupons.

**Problems**

1. **Read in the ‘coupons.csv’ file.**

We read the coupon.csv file and uploaded it to a Data frame. The Data frame contains 12,684 rows and 26 columns.

1. **Investigate the dataset for missing or problematic data:**

From simple data inspection, we have a data frame with 12,684 records of drivers that received a discount coupon for restaurants, bars, and coffee shops. However, before starting our coupon analysis, we observed that the data has some inconsistencies that need to be fixed. We will conduct a four-step cleansing.

**Step 1: Detect duplicates.**

We detected 74 row duplicates in the Data frame by using the **duplicated**() method. However, since we do not have a primary key for each row, we will assume that each record represents a different driver. Therefore, we will not eliminate identical rows. No changes are needed.



**Step 2: Resolve Structural Errors**

We inspected the data structure of the Data frame, and we observed a couple of numeric fields registered as String (object) fields. For instance, age and income fields.

We will not convert these fields to numeric because they contain data ranges, and we need to keep those ranges for coupon analysis. Therefore, no changes are needed.

A screen shot of a computer

Description automatically generated

**Step 3: Filter Outliers**

After inspection of the data, we noticed that we have some outliers in the Car column. For instance, we have a few drivers that “do not drive”. We will exclude them from the data frame because they are not drivers and therefore valid records. We exclude 22 rows.

**Step 4: Validate**

Since this data is descriptive, we do not have number columns to run statistics such as mean, media, std, skewness, etc. to validate the data behavior. However, we can review the unique values per each column and validate that they are reasonable.

Based on inspection, we observed that car column has “Car that is too old to install Onstar :D” value. We will maintain these rows and replace this car value with “Model Y”. Afterall, we do not focus on car models, but the coupon usage. We replaced 21 rows.

After all the replacements and exclusion, we have new a data frame called “coupon\_data\_cleaned” with 12,662 records. Please see below a summary table showing all the ins and outs from the original coupon data frame.

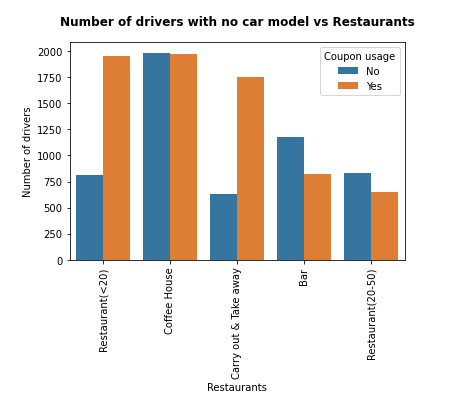
|  |  |  |  |
| --- | --- | --- | --- |
| Data frame name | Rows | Action | Value |
| Initial number of rows |  |  | 12,684 |
| Step 1: Duplicates | 74 | Do not remove | 0 |
| Step 2: Structural errors | 0 | Do not remove | 0 |
| Step 3: Filter Outliers | 22 | Remove | 22 |
| Step 4: Validate | 21 | Do not remove | 0 |
| Final number of rows |  |  | 12,662 |

1. **Decide what to do about your missing data -- drop, replace, other...**

To detect any empty values, we searched for all the unique values for each data frame column and printed the values for inspection. After inspecting the values, we observed the following fields with empty values.

* car
* Bar
* Coffeehouse
* CarryAway
* RestaurantLessThan20
* Restaurant20To50

For the car field, we expect to capture the car model data. However, we need to understand the meaning of the empty value. It could mean that we have no car for this driver, and he did not receive a coupon. Hence, to set some context, we prepared a “countplot” chart to identify whether drivers with empty car field used the coupon and went to a Restaurant, Bar, etc. From chart inspection, we observed that some of the drivers used the discount coupon.

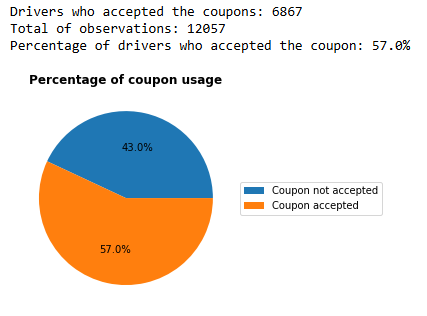


Therefore, they have a car, but the data does not show the car model. We will fill out the empty spaces with “Model X” value. So, they can be considered in the analysis. We replaced 12,576 rows.

For all the restaurant fields with empty values, we will drop all the rows that did not go to any of the restaurants (bars, coffee, etc.). There is no value in analyzing discount coupons if we do not know the driver consumer profile. After removing those rows, we retain 12,057 drivers.

1. **What proportion of the total observations chose to accept the coupon?**

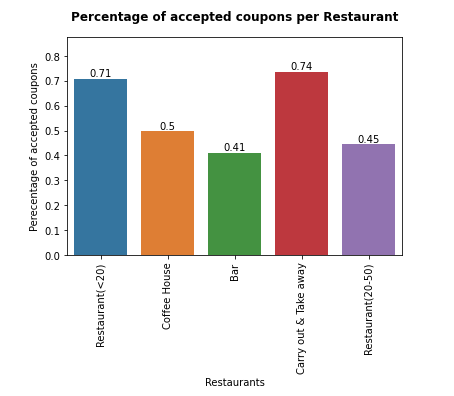
We filtered the data frame by the drivers who accepted the coupons, and then we compared it with the total number of observations. We found that 57% of the drivers accepted the coupons. Also, we prepared a pie chart to facilitate the inspection.



As we can see, the percentage of drivers who accept the coupon is superior to the percentage of drivers who reject the coupons. This suggests that distributing discount coupons is a good strategy to foster the consumption of food in restaurants.

1. **Use a bar plot to visualize the coupon column.**

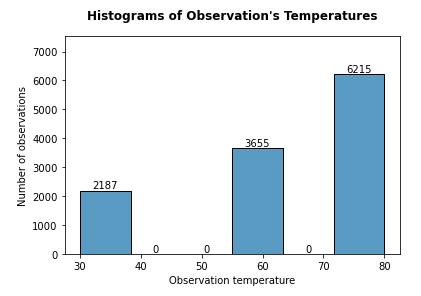
We prepared a bar plot for all the coupons, and we compared them against the coupon usage by restaurant. Per plot inspection, coupons is an effective strategy to foster food consumption in restaurants. No matter what type of restaurant you select, in all the cases, they have an acceptance rate superior to 40%. This strategy is especially effective for inexpensive restaurants and carry out restaurants, which have above 70% acceptance rate. For all the others, the acceptance rate is moderate, 40%.



At a glance, coupons is an effective strategy to foster food consumption in restaurants. No matter what type of restaurant you select, in all the cases they have acceptance superior to 40% of the cases.

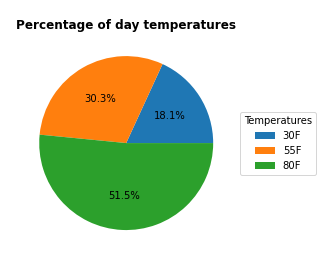
1. **Use a histogram to visualize the temperature column.**

We prepared a histogram to visualize the temperature distribution. Per histogram inspection, we observed that high temperatures are very common in the data frame.

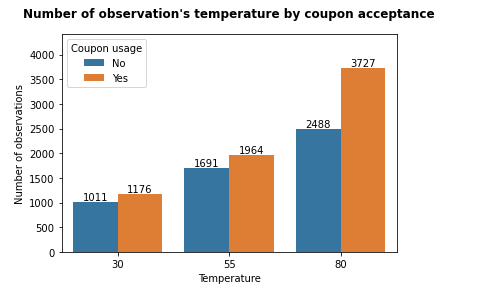


Also, we built a countplot and pie chart of the temperatures to have a better idea of the percentage distribution of temperature observations in the data frame, and, also, the coupon acceptance per temperature observation.

From the pie chart below, 80F is the most common temperature (more than 51% of observations), followed by the 55F temperature (30% of observations), and lastly 30F temperature (18% of observations).



From the count chart below, the data suggests that coupons are accepted for any observation disregarding the temperature. Further, it suggests that coupon acceptance increases with temperature, going from 1176 on 30F observations (53% of acceptance) to 3727 on 80F observations (60% of acceptance). On 80F observations, coupons are more accepted than any other temperature observation.



1. **Based on these observations, what do you hypothesize about drivers who accepted the bar coupons?**

For the coupon data, we observed that coupon acceptance rate is an effective strategy for all the restaurants (40% acceptance rate), especially for cheap restaurants and carry out restaurants (70% acceptance rate). Also, the coupon acceptance rate is higher on hot days (80F days), going from 1176 observations on 30F (53% acceptance rate) to 3727 observations on 80F (60% acceptance rate).

**Investigating the Bar Coupons**

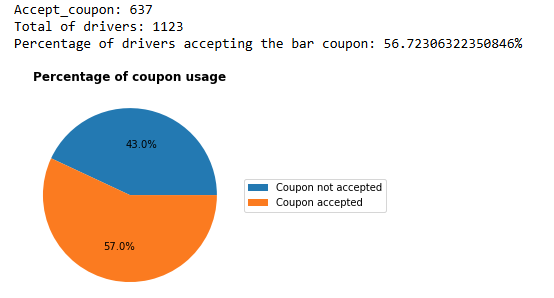
In this section, we will inspect the bar coupons to confirm whether they were used, and the acceptance rate for drivers who go three or fewer times a month.

1. **Create a new Data Frame that contains just the bar coupons.**

We filtered the previous data frame with cleaned data and excluded the observations that never go to the bar or with empty values. The Data frame contains 1,123 observations and twenty-six columns.

1. **What proportion of bar coupons were accepted?**

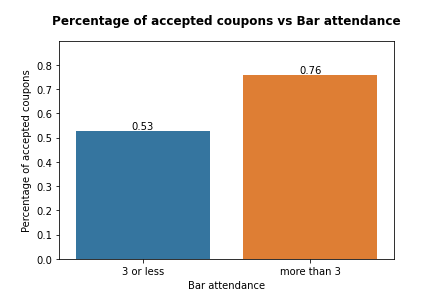
We count the total number of rows, and the total number of coupons that were accepted. After that, we conducted a simple division and obtained the percentage of drivers using a discount coupon (56.72%). Further, we prepared a pie chart for visualizations purposes, and we observed the same.



1. **Compare the acceptance rate between those who went to a bar 3 or fewer times a month to those who went more.**

We created a new column called “Bar\_less3” where we populated the column with ‘less3’ value if the “Bar” column had “less1” or "1~3", and ‘gt3’ value if the “Bar” column had “4~8” or "gt8". Once we classified all the drivers by “Bar\_less3” column, we prepared a barplot showing the two groups (‘3 or less, and ‘more than 4’). vs the “Accepted coupons”.

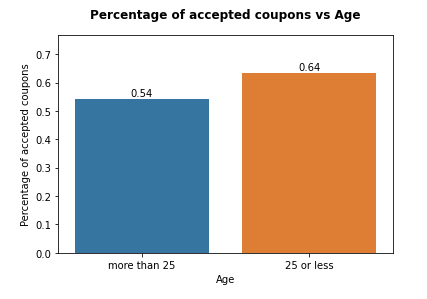
From the bar plot below, drivers who go to the bar 3 or fewer times have 53% of coupon acceptance, and drivers who go more than 3 times have 76% of coupon acceptance. Therefore, the more people go to a bar, the more likely they use the coupon discount.



1. **Compare the acceptance rate between drivers who go to a bar more than once a month and are over the age of 25 to all others. Is there a difference?**

Similarly, we created a new column called “Age\_less25” where we populated the column with ’25 or less’ value if the “age” column had “below21” or "21", and ‘more\_than\_25’ value if the “age” column had “26”, “31”, “36”, “41” or "46". Once we classified all the drivers by “Age\_less25” column, we prepared a barplot showing the two groups (‘less\_than\_25’, and ‘more\_than\_25’) vs the “Accepted coupons” column.

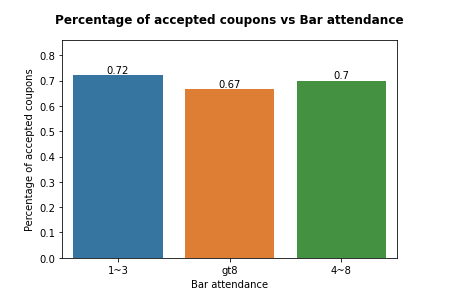
From the bar plot below, drivers who are 25 years old or younger have 64% of coupon acceptance, and drivers who are more than 25 years old have 54% of coupon acceptance. Therefore, the older the driver, the more likely he/she uses the bar coupon discount.



1. **Use the same process to compare the acceptance rate between drivers who go to bars more than once a month and had passengers that were not a kid and had occupations other than farming, fishing, or forestry.**

We applied filters to the data frame to the columns “Bar”, “Passanger”, and “occupation”. After that, we prepared a barplot showing Bar frequency groups (‘1~3’, ‘4~8’, and ‘gt8’) vs the “Accepted coupons” column.

From the chart below, all the drivers over who go more than 1 time to the bar with passengers that were not a kid and had occupations other than farming, fishing, or forestry have similar high acceptance rate (between 67% and 72%) to use the coupon.



1. **Compare the acceptance rates between those drivers who:**

* go to bars more than once a month, had passengers that were not a kid, and were not widowed *OR*

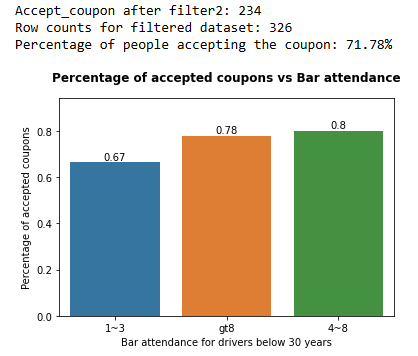
We applied filters to the data frame to the columns “Bar”, “Passanger”, and “maritalStatus”. After that, we count the total number of rows, and the total number of coupons that were accepted on the filtered data frame. Lastly, we conducted a simple division and obtained the percentage of drivers within the filtered dataset using a discount coupon (71.28%). We also, prepared a chart where we inspected the acceptance rate by bar attendance, and it shows a similar acceptance (between 67% and 72%) of coupons for each age group.

A graph with numbers and text

Description automatically generated with medium confidence

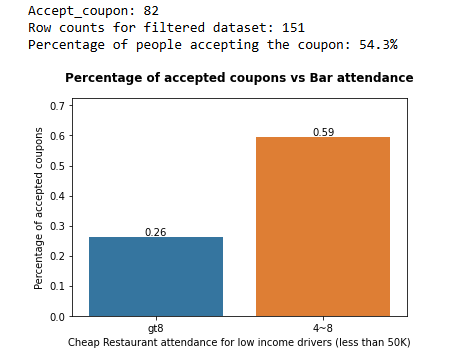
* go to bars more than once a month and are under the age of 30 *OR*

We applied filters to the data frame to the columns “Bar” and “age”. After that, we count the total number of rows, and the total number of coupons that were accepted on the filtered data frame. Lastly, we conducted a simple division and obtained the percentage of drivers within the filtered dataset using a discount coupon (71.78%). We also, prepared a chart where we inspected the acceptance rate by bar attendance for drivers below 30 years, and it shows that people that go more than 4 times to bar per month are more likely to use the coupons (nearly 80% of acceptance).



* go to cheap restaurants more than 4 times a month and income is less than 50K.

We applied filters to the data frame to columns “coupon”, “RestaurantLessThan20” and “income”. After that, we count the total number of rows, and the total number of coupons that were accepted on the filtered data frame. Lastly, we conducted a simple division and obtained the percentage of drivers within the filtered dataset using a discount coupon (54.3%). We, also, prepared a chart where we inspected the acceptance rate for bar tickets by cheap restaurant attendance for drivers who attend to cheap restaurants more than 4 times with an income less than 50K, and it shows that low-income people that go to cheap restaurants are more likely to use the coupons (nearly 80% of acceptance). Conversely, drivers who go more than 8 times to cheap restaurants will not accept the bar discount coupon (26% of acceptance).



1. **Based on these observations, what do you hypothesize about drivers who accepted the bar coupons?**

Based on prior charts, they are young people (less than 30 years) that go to the bar more than 1 time per month. They usually go when they are driving with friends or partners (no kids). In terms of their economy, they earn less than 50K and they go to eat to cheap restaurants between 4 and 8 times per month.

**Independent Investigation**

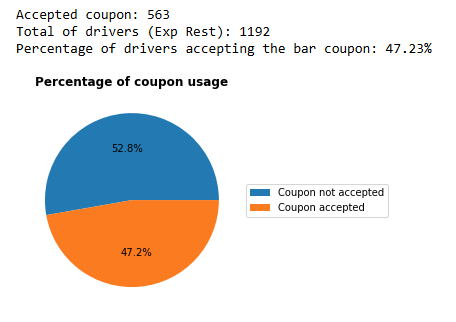
Using the bar coupon example as motivation, you are to explore one of the other coupon groups and try to determine the characteristics of passengers who accept the coupons.

1. **Create a new Data Frame that contains just the Expensive restaurant coupons.**

We filtered the previous data frame with cleaned data and excluded the observations that never go to the bar or with empty values. The Data frame contains 1,192 observations and twenty-six columns.

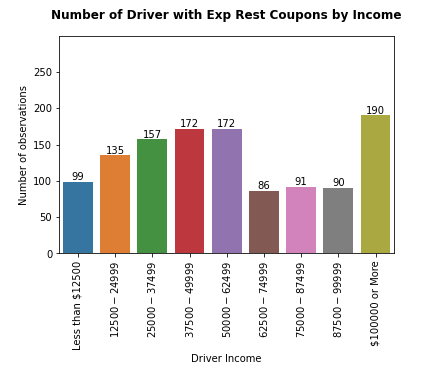
1. **What proportion of expensive restaurant coupons were accepted?**

We count the total number of rows, and the total number of accepted coupons. After that, we conducted a simple division and obtained the percentage of drivers using a discount coupon (47.23%). Further, we prepared a pie chart for visualizations purposes, and we observed the same.



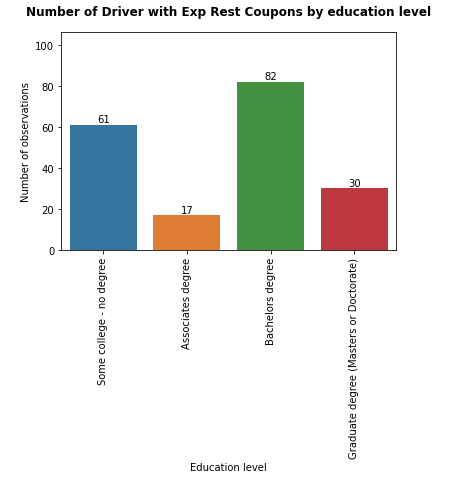
1. **What is the income distribution of the drivers with expensive restaurant coupons?**

First, we will classify the drivers with Exp Restaurant coupons by income and prepare a count plot. From inspection of the count plot, we observe that the biggest number of observations comes from the drivers who earn more than 100K per year.



1. **What is the education level of the drivers earning more than 100K with expensive restaurant coupons?**

We will inspect the 190 drivers with Exp Rest coupons earning more than 100K per year, and we will classify them by education degree. We prepared a count plot, and observed 82 drivers with Bachelor degree, which represents 43.15%. Therefore, we will inspect the Bachelor group.



**Note**: We observed that drivers earning more than 100K with Associate degree do not eat often in expensive restaurants (17 observances) compared to any other driver group earning more than 100K (some college, bachelor, graduate degree).

1. **What is the expensive restaurant attendance and coupon usage for drivers earning more than 100K with bachelor’s degree and expensive restaurant coupons?**

We will inspect the 82 drivers with Bachelor education and Exp Rest coupons earning more than 100K per year, and we will classify them by attendance. We prepared a count plot by attendance and coupon acceptance and observed that as the frequency of attendance increases (less1, 1~3,4~8), the attendance values decrease (43, 35, 4). Further, the more monthly visits per group increase (43, 35, 4), the coupon acceptance numbers increase (15, 18, 3) as well. It is important to note that coupons will be more likely to be accepted for any driver that goes more than 1 time per month (18>17 and 3>1) to expensive restaurants.

A graph of a bar graph

Description automatically generated

1. **Based on these observations, what do you hypothesize about drivers who accepted the expensive restaurant coupons?**

In conclusion, expensive restaurant coupons will be more likely accepted by drivers earning more than 100k with a bachelor’s degree when they go more than 1 time per month to an expensive restaurant. Additionally, we observed that drivers earning more than 100K with Associate degree do not eat often in expensive restaurants (17 observances) compared to any other driver group earning more than 100K (some college, bachelor, graduate degree).