Practical Exam - Coffee Shops

Instructions

- Use Python or R to perform the tasks required.
- Write your solutions in the workspace provided from your certification page.
- Include all of the visualizations you create to complete the tasks.
- Visualizations must be visible in the published version of the workspace. Links to external visualizations will not be accepted.
- You do not need to include code unless the question says you must.
- You must pass all criteria to pass this exam. The full criteria can be found here.

Background

Java June is a company that owns coffee shops in a number of locations in Europe.

The company knows that stores with more reviews typically get more new customers. This is because new customers consider the number of reviews when picking between two shops.

They want to get more insight into what leads to more reviews.

They are also interested in whether there is a link between the number of reviews and rating.

They want a report to answer these questions.

Data

The dataset can be downloaded from here.

Column Name	Criteria
Region	Nominal. Where the store is located. One of 10 possible regions (A to J). Missing values should be replaced with “Unknown”.
Place name	Nominal. The name of the store. Missing values should be replaced with “Unknown”.
Place type	Nominal. The type of coffee shop. One of “Coffee shop”, “Cafe”, “Espresso bar”, and “Others” Missing values should be replaced with “Unknown”.
Rating	Ordinal. Average rating of the store from reviews. On a 5 point scale. Missing values should be replaced with 0.
Reviews	Nominal. The number of reviews given to the store. Missing values should be replaced with the overall median number.
Price	Ordinal. The price range of products in the store. One of “$”, “$$” or “$$$” Missing values should be replaced with ”Unknown”.
Delivery Option	Nominal. If delivery is available. Either True or False Missing values should be replaced with False.
Dine in Option	Nominal. If dine in is available. Either True or False Missing values should be replaced with False.
Takeaway Option	Nominal. If take away is available. Either True or False Missing values should be replaced with False.

Tasks

Submit your answers directly in the workspace provided.

1. For every column in the data:
    a. State whether the values match the description given in the table above.
    b. State the number of missing values in the column
    c. Describe what you did to make values match the description if they did not match.
2. Create a visualization that shows how many stores were given each rating. Use the visualization to:
    a. State which category has the most number of observations
    b. Explain whether the observations are balanced across categories
3. Describe the distribution of the number of reviews. Your answer must include a visualization that shows the distribution.
4. Describe the relationship between number of reviews and rating. Your answer must include a visualization to demonstrate the relationship.
5. The business wants to predict the number of reviews a store will get using the data provided. State the type of machine learning problem that this is (regression/classification/clustering).
6. Fit a baseline model to predict the number of reviews a store will get using the data provided. You must include your code.
7. Fit a comparison model to predict the number of reviews a store will get using the data provided. You must include your code.
8. Explain why you chose the two models used in parts 6 and 7.
9. Compare the performance of the two models used in parts 6 and 7, using any method suitable for the type of model. You must include your code.
10. Explain which model performs better and why.

Sample Solution

You can find an example solution for this sample here.