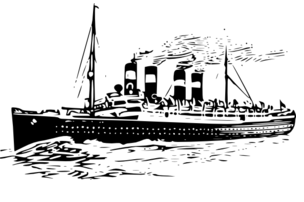
# Lab 2 Regression

Please submit your answers in this word document, including detailed discussion for each question with supportive screenshots of your analysis in python. Be sure your answers are numbered to correspond to the directions so that your responses are clear.

## Linear Regression

Welcome to CastAway Cruise Lines ™!

CastAway Cruise Lines™ is a short to midterm cruise operator based out of the port of Miami. We maintain a fleet of 3 elegant vessels that each have approximately 750 cabins per ship. Each ship has 30 suites (S), 120 Balcony (B) cabins, 150 cabins equipped with a window (W) and 450 Interior (I) cabins. In addition to room type, we track total cost, room costs, ship board expenses, casino expenses, and excursion expenses for all of our customers. Our primary concern is maintaining our customer satisfaction amongst our patrons. For all of analysis, make sure that Customer Satisfaction Average is ‘Target’ while Overall\_1\_5\_Satisfaction is ‘Rejected’.

As our newest (and only) data analyst you are tasked with the following projects:

1. Is there a relationship between how much a patron’s entire trip costs and their satisfaction? Provide the visual and numerical values that you referenced and justify your conclusion. (Hint: Calculate the correlation between the variables).

**Answer:** Yes, there is certainly a relationship between the Total Cost and their satisfaction.

**Correlation:** The positive correlation of **0.4683** suggests that the higher cost would be associated with higher satisfaction. The moderate value of correlation implies that there is some association, but it is not definitive, there may be some unaccounted factors which impact satisfaction.

**Linearity:** The scatter plot does not show strong linear pattern which indicates variability of customer’s approach towards value and satisfaction with respect to cost.

A screen shot of a graph

Description automatically generated

1. Build the best possible model you can, without any categorical variables, to predict customer satisfaction (Make sure the role of Customer Satisfaction is your dependent variable). Show the screenshot of the output of the parameter estimates and significance. How predictive is this model? Is there any potential for multicollinearity? How did you come to this conclusion?

**Answer:**

**p-value:** The p-value suggests that the model is not statistically significant. It does not provide a significant prediction of the dependent variable.

**Misclassification Rate:** 0.4577777777777778 is a high value for the misclassification rate. This implies that the model’s predictiveness is not much accurate.

**Pseudo R-squared:** 3.690e-11 is a very low value which point towards the model having low predictive ability.

**Multicollinearity:** The Variance Inflation Factors (VIF) is a measure that quantifies the extent of multicollinearity in a least square regression analysis. The high values for VIF for most of the features suggest that there is a strong multicollinearity among the predictors in the model.

The below screenshot displays the code output.

A screenshot of a computer

Description automatically generated

1. Does the choice of the type of room have significant impact on the customer satisfaction if it is the only predictor? If so which room type(s) is different from the reference type (cabins with a window)? How did you come to this conclusion using the software output?

**Answer:**

**F-statistic and p-value:** Since the p-value for the F test is less than 0.05, we can state that the model is statistically significant.

**F-**statistic: 1.995

**p-**value: 0.0326

**Individual Room Type:** the p-value for each individual room type is not less than 0.05 for any room type. This implies that individually none of the room type have statistically significant difference in customer satisfaction compared to reference type (cabins with a window).

**Overall Significance:** As the p-value is less than 0.05, this signifies that the room type as a group is statistically significant and affects overall customer satisfaction. But individually, the room type does not impact the customer satisfaction.

The below screenshot displays the code output.

A screenshot of a computer

Description automatically generated

1. Considering all predictor candidates including the categorical variable(s), create the best model you can and show the screenshot of the output of the parameter estimates and significance. Also, without real implementations, please discuss a general idea of how to improve this model by creating or adding other categorical variables.

**Answer:**

**Adding other categorical variables to improve model’s performance:** Categorical variables capture the details that the numerical or continues variables may not efficiently capture.

**Example:** Adding “customer segment” can give us insights about preference of business travelers than the leisure travelers.

**Flexibility**: Adding more categorical variables increases **flexibility** of the model. Flexibility makes the model adapt better to real world situations.

**Bias:** Adding more categorical variables will reduce the **bias** in the model.

**Example**: Having “age group” in the model will not make generalized assumptions for all age groups.

The below screenshot displays the code output for the model created with all predictor variables including categorical variables.

A screenshot of a computer

Description automatically generated

1. How can you use this model to increase customer satisfaction? Give specific examples and tie it back to the individual coefficients.

**Answer:**

We can analyze the individual coefficients to increase customer satisfaction with respect to this model.

* **R-squared:** 0.338

This indicates that the model has about 33.8% variance in customer satisfaction. This is a good figure to start with identification of actionable areas.

* **Total Cost:**

**Coefficient: 0.0080, p-value: 0.180**

The positive coefficient suggests that higher price is associated with higher satisfaction.

**Action:** We can focus on providing exceptional value for higher costs.

* **Room Costs :**

**Coefficient: -0.0014, p-value: 0.047**

It is a significant value with negative impact which implies that higher costs may lead to lower satisfaction.

**Action:** We can adjust the pricing for the rooms and align it with the customer expectations.

* **Excursion Expenses**

**Coefficient: -0.0050, p-value: 0.409**

Non-Significant negative trend which means higher excursion cost result in lower satisfaction.

**Action:** Adjust the pricing of excursion tours to provide value and align it with customer expectations.

**A screenshot of a computer

Description automatically generated**

## Logistic Regression

1. Using the same dataset as you did for the Linear Regression portion of the assignment, please estimate a logistic regression to predict whether a customer booked a suite (Room Type = S) as opposed to any other type of room. (HINT: You need to create a new dependent variable which =1 if Room Type = S and 0 otherwise) Select 2 independent variables which you believe to be most important and explain why you chose those variables in your model.

**Answer:**

The independent variables selected are “Room Costs” and “Overall 1-5 Satisfaction”, these were important to predict if a customer booked a suite or not.

Room Cost indicates the type of room booked. If a suite is booked it would be more expensive. Higher satisfaction would be directly related to booking higher quality rooms and having a comfortable stay. Excursion Expenses

The below screenshot shows the code output and the misclassification rate.

A screenshot of a computer

Description automatically generated

1. Add 2 additional variables. How did your misclassification rate change? Which model do you believe is better and why? Please be sure your results screenshot includes anything you reference in your answer.

**Answer:**

**Additional Variables: Casino Expenses and Excursion Expenses**

Both the models are perfect models with a misclassification rate as 0.0. In this case the model came out to be a perfect model otherwise the a simpler model is a better model. In our case, the first model with two independent variables is simpler and preferable.

Added features would help to capture more complex relationships and could perform better on varied, unseen data.

The below screenshot shows the code output and the misclassification rate.

A screenshot of a computer program

Description automatically generated