Assignment-based Subjective Questions

# Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

# Total Marks: 3 marks (Do not edit)

# Answer: <Your answer for Question 1 goes below this line> (Do not edit)

# #The season with least booking was spring

# #There are no much variations observed in the count of booking with regards to different days of the week

# #clear days with no or least cloud has highest booking

# #it is observed that 2019 has the highest bookings

# 

**Question 2.** Why is it important to use **drop\_first=True** during dummy variable creation? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

# Answer: <Your answer for Question 2 goes below this line> (Do not edit)

# It is used to remove any extra columns while creating dummy variables

# It is used to reduce collinearity

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

**Total Marks:** 1 mark (Do not edit)

# Answer: <Your answer for Question 3 goes below this line> (Do not edit)

# Temp and atemp has highest correlation with the target variable which is count

**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

# Answer: <Your answer for Question 4 goes below this line> (Do not edit)

# The values with low vif was iterated and chosen for further filtering.

# The normal distribution of error terms was taken into account.

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

# Answer: <Your answer for Question 5 goes below this line> (Do not edit)

# Temperature

# Non-rainy pleasant climate and the month of sept-oct

# General Subjective Questions

**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)

**Total Marks:** 4 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 6 goes here>

# The linear regression algorithm goes in this way. Finding the summary for the p values off the variables, finding the vif and the ruling out the values having high vif and trending p values to then shortlist and iterate the process till the number of significant variables are small.

**Question 7.** Explain the Anscombe’s quartet in detail. (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 7 goes here>

# It is a set of 4 data sets that has identical statistics like mean stdev etc but upon plotting appear to have vast difference while plotted

**Question 8.** What is Pearson’s R? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 8 goes here>

# This coefficient provides the correlation between the two sets of data

**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 9 goes here>

# In order to normalize the data the independent variables within a particular range are trained this is called scaling. Normally data collected will not be of the same scale .so modeling them is= such a fashion that they all come under same category is important for uniform and speedy calculations.

# Normalized scaling brings the values between 0 and 1 while standardization does another was by replacing it by the z value

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 10 goes here>

# It happens when the multicollinearity between variables is soo perfect in a regression model

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 11 goes here

# Graphical tool used t compare the distribution of dataset to the theoretical distribution.