# 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)

# Based on the analysis I have done, categorical variables have a significant influence on the dependent variable (“cnt”) i.e. bike rental demand. Below are the highlights of few such categorical variables:

# Year (“year”): Rentals increased significantly from 2018 to 2019, as indicated by the positive coefficient of “year”. This could reflect increased demand over time, potentially due to growing popularity, expanded bike-sharing infrastructure, or marketing efforts.

# Season (“season”): “Summer” and “Fall” seasons showed increase in demand whereas “Spring” and “Winter” demonstrated a decline

1. **Weather Situation(“weathersit”):** Misty or cloudy days (“weathersit” category level 2) reduced rentals compared to clear days (“weathersit” category level 1). Bad weather conditions (light snow or rain i.e. “weathersit” category level 3) further reduced rentals.
2. Then there are some minor impacts month (“Sep”) i.e. September month or Fall showed increase in demand

# 

**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)

# For categorical variables with n levels, we only need to create dummy variables for n-1 levels as we can explain the left-out variable with the values of other variables. This also helps avoid multicollinearity issue. Multicollinearity caused by redundant dummy variables can lead to instability in the estimated coefficients, causing large variances and making the model sensitive to small changes in data. Dropping one dummy variable helps ensure that each feature contributes uniquely, making the model’s estimates more stable and reliable.

**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)

# Temperature (“temp”) has the highest correlation with the target variable of bike rental demand (“cnt”). The Pearson’s r correlation value is 0.63. I dropped the Temperature Feel variable (“ atemp”) due to multicollinearity issue (with Temperature (“temp”) variable) which also showed a similar correlation.

**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)

I used the following methods to validate the assumptions of Linear Regression:

* **Linearity**:
  + I checked for linear relationships between predictors/independent variables and the target variable (“cnt”) during Exploratory Data Analysis (EDA) in section 6.2 of the Jupyter notebook “bike-sharing-model-abhishek-mukherjee.ipynb” using a **correlation heatmap** and **scatter plots (“temp” and ”atemp” variables showed promising linear relationship)**. Linear relationships support the assumption that predictors have a linear association with the target.
* **Normal Distribution of Residuals**:
  + After fitting the model, I plotted a seaborn distplot of the residuals (differences between actual and predicted values) in section 12 of the Jupyter notebook “bike-sharing-model-abhishek-mukherjee.ipynb”. The residuals/error terms showed a normal distribution.
* **Homoscedasticity**:
  + I created a **scatter plot of residuals vs. predicted values**  in section 12 of the Jupyter notebook “bike-sharing-model-abhishek-mukherjee.ipynb”. For homoscedasticity, the residuals should show constant variance across predicted values, with no clear pattern. The residual plot did not show any systematic pattern, indicating that the assumption of homoscedasticity was met.
* **Multicollinearity**:
  + I assessed multicollinearity before modelling for the “temp” and “atemp” variables which showed similar patterns and were highly correlated with each other in the heatmap visual (correlation of 0.99) in EDA section 6.2 of the Jupyter notebook. Then I used **Variance Inflation Factor (VIF)** after the initial model fitting. I also dropped the “const” variable which had a high value of ~9.87 indicating multicollinearity issues.
* **Independence of Errors**:
  + The **Durbin-Watson statistic** (reported in the OLS summary) was close to 2, suggesting that residuals are uncorrelated (no significant autocorrelation), which is ideal for linear regression.

**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)

# Based on my final model (model 6 in manual feature selection post RFE) which had six variables below are the top 3 predictors/features:

# Spring Season: The variable "spring" has a coefficient/weightage of -0.8644. The bike demand will significantly fall in spring and the company needs to invest in some marketing strategies to attract more people during spring. "Summer" and "Fall" seasons drives demands up.

# Weather Conditions: Weathersit variables like "Light Snow" (coefficient/weightage of -1.4208) and "Mist" (coefficient/weightage of -0.4554) drives demands down.

# Year: This variable "year" has a coefficient/weightage of 0.9081. Year on year demand is increasing. Covid-19 may have temporarily stifled demand but as the pandemic is gone now (or is less serious) the demand is expected to rise again. People are also becoming more climate conscious and prefer bikes for shorter transits and for health benefits.

# 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>

**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>

**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>

**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>

**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>

**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>