# 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)

- cnt seems to be impacted by season

- number of rented bikes has increased from 2018 to 2019
- there is relation of cnt with month
- people seem to rent more bikes when it's not a holiday
- more bikes are rented in clear weather

---

**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)

Say there are "n" unique values in a categorical column then get_dummies function creates n columns. But we need only (n-1) columns to deduce all possible values.
When we use drop_first=True it drop first dummy column leaving (n-1) columns in the dataset.

---

**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**

---

**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)

- There are certain assumptions in linear regression. Residual analysis is done to ensure that error terms are meeting those conditions.
- Assumptions are : Error terms should be normally distributed. We check this by plotting a distplot of residues. Mean of this distribution should be 0. Sum of all residues should be 0 as well,
- We also see distribution of predicted values vs actual values. If it is distributed around a line at 45 degree with no clear pattern and no outliers. It means that residues have constant variance.

---

**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)

1. Temp: temperature
2. Hum: humidity
3. yr: year

## 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>

Step 1: Reading, understanding and visualizing data

- Data is read using pandas library into a dataframe
- To understand the data data dictionary is used.
- Check for missing or null values, fix if there are any datatype mismatches
- Divide the columns into categorical and numerical
- Plot pair plots for numerical data. See how these relate to target variable.
- Plot box plot for categorical columns vs target variable. See if there is any pattern.

Step 2: Data Preparation

- Convert values of binary columns to 0/1
- Create dummy variables for multivalued categorical columns
- Split the data into train and test
- Rescale the features using normalization or standardization. Transform and fit the training data using sklearn scaler but only transform the test data using the scales the scaler has learnt from training data.

Step 3: Training the model

- Feature selection: If number of predictors are more (>10) then use automated (eg RFE) and manual approach of feature selection otherwise forward/backward/stepwise method can be used. Other advanced techniques ( like lasso) are also available .
- Train the model using statsmodels.api as it gives detailed model stats.
- Check if coefficients are in reasonable range
- Check if p-values are less than 0.05. If not then these features need to be dropped and model needs to be retrained.
- Check if VIF is less than 5. If not then it indicates the feature is correlated to other features. It needs to be dropped and model needs to be retrained.
- Check if r-square value is greater than 0.5. It means that overall model is able to explain more than half the variance in dataset. So it is acceptable model.

Step 4: Residual Analysis

- There are certain assumptions in linear regression. Residual analysis is done to ensure that error terms are meeting those conditions.
- Assumptions are : Error terms should be normally distributed. We check this by plotting a distplot of residues. Mean of this distribution should be 0. Sum of all residues should be 0 as well,
- We also see distribution of predicted values vs actual values. If it is distributed around a

line at 45 degree with no clear pattern and no outliers. It means that residues have constant variance.

- 

Step 5: Model Evaluation

- We check the r-square score of test data. If its more than 0.5 that means model is predicting with good accuracy.

Step 6: Drawing inferences from model

- From model summary we know the coefficients of predictors. So our model is b0 + b1*x1 + b2*x2 …..
- Features with higher coefficients are more significant contributors in the growth of target variable
- Some coefficients may be negative which indicate inverse relationship.

**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>

**Anscombe's quartet is a set of 4 datasets which have very similar summary statistics but are very different when plotted. It emphasizes on the need of plotting the dataset and not entirely depend on statistics.**

**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>

Pearson's R is a measure of strength of linear correlation between 2 variables. Its values is in the range [-1,+1]. Positive values means positive correlation and negative value means inverse correlation. Values closer to 1 or -1 means stronger correlation. Pearson's R assume linear relationship if data is non-linear then its value is misleading. It is also sensitive to outliers.

**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>

**Scaling involves transforming numerical values in a dataset so that all columns have comparable ranges while preserving the relative relationships between values. In a dataset, different columns may have varying units—some values might be in thousands, while others in decimals. This variation can lead to model coefficients with a wide range, increasing model complexity and slowing down gradient descent during training. Scaling helps standardize these values, enhancing model efficiency and training speed.**

**Normalized scaling formula : ( x – xmin ) / (xmax – xmin )**

- **It brings all values in the range [0,1]**
- **Categorical values with 0/1 discreet values are not impacted**
- **It takes care of outliers.**
- **This method is generally used for scaling in model training when we need bounded values.**

**Standardised scaling formula : ( x – mu ) / sigma where mu is mean and sigma is variance of data.**

- **It changes the data in a way that mean become zero and variance is 1**
- **Some data points including outliers can go beyond +-1. 68% data converges between [-1,+1].**
- **It is used when we need normal distribution.**

**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 means that predictor can be perfectly explained by using other predictors in the 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>

**A Q-Q plot is a graphical way to see that data follows which distribution ( normal, uniform or something else). It's done by plotting the quantiles of data vs quantiles of distribution. If data belongs to that distribution then the points will lie on a straight line.**

---