# 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)

From my bivariate analysis of categorical variable it is evident that rental bike increased in year 2019, customers are renting more bikes between May-October means these have more clear sky days and very less rain season as it is also evident from 'weathersit' vs 'cnt' analysis, also bike are more rented when sky is clear, or partly cloudy.

Also more people are renting bike on holidays and weekends as compare to week-days, these are casual people how are renting for maybe roaming around and for fun but not for daily commute.

---

**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)

Using this command remove extra dummy variable column and also remove original column which help in model building, because those column are redundant column and does not convey any extra information

---

**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)

**Registered** variable has highest correlation with the target variable.

---

**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)

After building the Linear regression Model on the training set we need to calculate the residual analysis. Using formula (y_train - y_train_predict), plot distribution plot to check the distribution, if this comes as normal distribution means graph peaked at 0 Ideally. This shows all the error / residual datapoints are evenly distributed and there difference equates to 0.

---

**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 the final model, Temparature with coef=4727.7709, year with coef=1960.6886, season_4 with coef=1172.7494 are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

---

## 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>
Linear regression is supervised machine- learning algorithm, it learn from defined datasets. In this algorithm learns from independent variables and make predictions on the target variables. These predictions are based on interpolation. Interpolation means predictions are based in between knowns dataset points, it cannot be used for extrapolation.Linear regression algorithm work with some assumptions like Linearity, Homoscedasticity, Independent X and Y pairs for training and testing, Linear regression error have to follow Normal Distribution. Since it is supervised Machine learning Technique there should be no hidden values or missing data in the dataset.
Linear regression follow straight Line equation which is $Y=mx + C$, or $Y= \beta0 + \beta1x$. where $\beta0$ is intercept and $\beta1$ is slope of gradient.

---

**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 method of checking the dataset is same or not. It consists of 4 dataset these have identical statistical properties like all have same mean, standard deviation, r-square, correlation. But they have different scatter plot on graph means data representation is different for all 4 dataset. Anscombe's shows the importance of EDA before model building and drawbacks on depending too much on statistical summary. Anscombe's quartet help to spot outliers, trends and other deviations that get neutralized in summary 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>
It is the correlation coefficient between 2 variables X & Y which measure the distance of datapoints from best fit linear regression line. Fundamentally it draws the best fit line using datapoints so that all data points have least distance from the line. Pearson's R can take range from +1 to -1. 0 value means there is not association between datapoints of 2 variables. R=+ve means Positive correlation means value of X & Y increases together, r=-ve means negative correlation, where X increases but Y decreases.

---

**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 is the pre-process which we do before model building, it affects the coefficient only for example lets take 2 predictors from the dataset , numerical value of temperature and humidity are both different. So we perform certain methods so that both values become comparable and does not create false model. 2 types of scaling

1. Standardized scaling:- In this all all dataset get scaled as normal distribution with mean 0 and standard deviation
2. Normalised Scaling:- In this dataset get converted into range of 0 and 1. All predictors columns values comes between range of 0 and 1.

---

**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>

I have seen in my model some variables have VIFs values as '**inf' .** This means **Perfect Multicollinearity** means one predictors is an exact linear combination of other. One variable can be perfectly explain by other variable. For example temp and atemp has high correlation this create perfect collinearity. This is very bad for model building as this create overfitting and wrong prediction for test variables

---

**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>

Q-Q plot or quantile-quantile plot is graphical method to determine the probability distribution, Q-Q plot generally used to check the normal distribution graph.

Quantile-Quantile plot help graphically to check the assumption of linear regression that residual follows the normal distribution. This plot compare the residual plot with theoretical quantile of normal distribution. Deviation are easily detected if there is skewness or outlier detection if graph has heavy tails. Early usage of this plot help detect if linear regression will work on the dataset or we need some other model.

---