**Presented by**

**Vibhudendra S**

**upGrad & IIITB | Data Science Program - February 2024**

**Batch ID 5702**

**Dt. 15.10.2024**

**Assignment-based Subjective Questions**

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

*Ans: In the raw dataset, we only had one variable as categorical i.e., 'dteday' which we dropped as it is not a time series based business problem. We also converted two more variables as categorical based on the business problem i.e., 'season' & 'weathersit'. We understood the following*

- *Majority demand for shared bikes is clearly around 6,000 for all 3 seasons(Summer, fall & winter) except during spring whose demand is around 4,000 only, 33% lesser*
- *Majority demand for shared bikes is clearly around 6,000 in misty & clear weather conditions but dips to around 3,000 only during light snow weather, 50% lesser*

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

*Ans: The remaining dummies nil values indicates the presence of dropped dummy variable. Explained using the assignment example as under*

| season_spring | season_summer | season_winter |
|---|---|---|
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |

*The dropped dummy variable here is season_fall which means that we have 0 values in all these dummy variables(season_spring, season_summer & season_winter).*

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

*Ans: **'registered'** variable has the highest correlation with target variable.*

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: *Validated assumptions of Linear Regression as under*

- *R-Squared value & Adjusted R-Squared value: Used RFE method to build the model on train dataset. At every model fit, we assessed R-square score & Adjusted R-Squared value so that we avoided overfitting the model(first 4 iterations had both of these scores as perfect 1 which meant overfitting) & maximized as General Linear models with final scores of 0.825 & 0.822 respectively.*
- *p-value: eliminated those variables having p-values greater than 0.01*
- *VIF score: eliminated those variables having VIF greater than 5*
- *Distribution of error terms had normal distribution with mean centred around Zero.*

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

*Ans: **casual users, weather & season** are the top 3 features ontributing significantly towards explaining the demand of the shared bikes*

*Co-efficients are 0.486 * casual - 0.2285 * weathersit_light snow - 0.2124*season_spring*

---

**General Subjective Questions**

1. Explain the linear regression algorithm in detail. (4 marks)

*Ans: predicting the value of continuous target variable using linear relationship of independent variables having the below assumptions is defined as Linear Regression algorithm.*

*Key Assumptions are*

a. *Linear relationship between Target variable & independent variable.*
b. *Error terms are normally distributed with mean centred around Zero.*
c. *Independence of error terms*
d. *Constant variance in errors*

*Ordinary Least squares(OLS) methodology is used to fit the best linear relationship/best fit line between target & independent variables.*

2. Explain the Anscombe's quartet in detail. (3 marks)

3. What is Pearson's R? (3 marks)

Ans: *It's a measure of linear relationship between 2 variables with -1(highest negative correlation) & 1(highest positive correlation), 0(No correlation).*

3. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

*Ans: Scaling is a transformation technique used to bring all features on a similar measure scale to determine the co-efficients in Generalized Linear models. If we don't perform scaling, then the model predicts abnormal co-efficients to certain features only based on its actual data values irrespective of its relevance to target variable. For ex: If Area of a house is in 1,500 sqft & no. of bathrooms are 4 then the data values 1,500 & 4 in as is condition logically have abnormal co-efficient to area than bathroom. The Price in this case being the target variable doesn't relate to so much abnormality in its respective co-efficient without scaling methodology. The main difference between standardized scaling & normalized scaling is that normalized scaling has values proportioned/scaled between 0 & 1 whereas Standardization transforms data to have a mean of 0 and a standard deviation of 1.*

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

*Ans: if there is a perfect correlation(-1 or +1) between two variables, then VIF will be infinite*

*VIF = 1/(1-r2) where r2 is a measure of correlation falling between -1 & 1.*

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

*Ans: Q-Q plot is a quantile quantile plot, a graphical representation of data densities between actual data values & predicted data values. This is mainly used to conclude these data sets have a common distribution in linear pattern.*