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 the given dataset, the categorical columns are season, year, month, holiday, weekday, working day and weather situation. The inference of their effect on dependent variable 'cnt' are,

1. **Season**: *Fall* records the higher rental bookings followed by *summer* and *winter*. Spring records the very less bike rental bookings
2. **Year**: We have 2018 and 2019 as years. The total bike rental bookings are increased from 2018 to 2019.
3. **Month**: There is upward trend from Jan to June and gradually decreasing. September marks the highest rental bookings among all the months.
4. **Holiday**: There is very low mark captured during holidays. Also, very high booking I also captured on Holiday itself.
5. **Weekday**: There is no major difference we could capture here. But 0 and 6 marks high stating that weekends capture more bike rentals
6. **Working Day**: There is no significant variations we could see here.
7. **Weather situation**: When there is cleary sky, the total bike rental is very high followed by Mist. During LightSnow, the bookings of rental bike is significantly low.

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)

The rule for dummy variable creation is '**n-1**'. Because, to avoid the issues related to multicollinearity and interpretation from redundant variable. If there are 'n' categories in the variable, we should create 'n-1' dummy variable. In python, by default it creates 'n' dummy variable. Due to this, we have to use 'drop_first = True' which drops the 1st column to adhere to the rule of 'n-1' dummy variable creation. By doing so, we can avoid multicollinearity, avoid misinterpretation, reduce the complexity of the model.

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)

The temperature 'temp' has the highest correlation with target variable 'cnt'

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)

1. **Residual Analysis**: Calculating, 'y_train - y_train_pred', we can get the residuals. By plotting it in the histogram, pattern should follow the normal distribution with mean 0. By this, we can confirm that the errors terms are normally distributed
2. **Multicollinearity**: By calculating VIF (Variance Inflation Factor), if VIF<5, the multicollinearity is very less and the variable has VIF>5, the variable will be eliminated. Hence, the finalized variable should have VIF<5.
3. **Homoscedasticity**: Error terms are independent to each other. We need to plot the scatter plot for the error terms. It should not follow any pattern.
4. **Linear pattern**: Plot the predicted values with the actual value in a scatter plot. We should see the linear pattern.

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)

In my model, significant top 3 features are

1. **Yr** : Coeff (1.0489), P-value(0.000)
2. **Temp**: Coeff(0.438), P-value (0.000)
3. **season_spring :** Coeff(-0.4727), P-value (0.000)

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 type of predictive analytics comes under supervised machine learning algorithm which can be applied on continuous numerical variable. It is used to estimate the linear relationship between variables (i.e) linear relationship between independent variable (predictor) and dependent variable (target). Since this regression algorithm is linear, it adopt the equation of straight line to predict the relationship.

$\mathbf{Y=mx+c}$ ----→ Equation of a straight line

m- co-efficient of 'X' (Slope)
c – Intercept (Constant)

This will explain the change in 'Y' for every change in 'X' with some magnitude. This algorithm will find the best fit line in order to reduce the error. There are two types of linear regression,

1 **Simple Linear Regression**: To identify the relationship between target/dependent variable and only one independent (predictor) variable. The equation of SLR is,

$\mathbf{Y = \beta_0 + \beta_1 X_1 + \varepsilon}$ -→ SLR equation

2 **Multiple Linear Regression**: To identify the relationship between target/dependent variable and multiple independent (predictor) variable. The equation of MLR is,

$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots \beta_n X_n + \varepsilon}$ -→ MLR equation

β_0 – Intercept (Constant)
β_1 , β_2 ,β_3 , …… β_n - Co-efficient of predictor variable
X_0, X_1, X_2 ….. X_n – Predictor / independent variable
Y – Target variables

Both the models follow specific set of assumptions. While performing the linear regression, the model made those assumption true.

1. There should be a linear relationship between target and predictor variable
2. Error terms are normally distributed with mean 0
3. Error terms are independent to each other
4. Error terms should have constant variance (Homoscedasticity)
5. Adding more variable – Overfitting and Multicollinearity
6. Feature selection

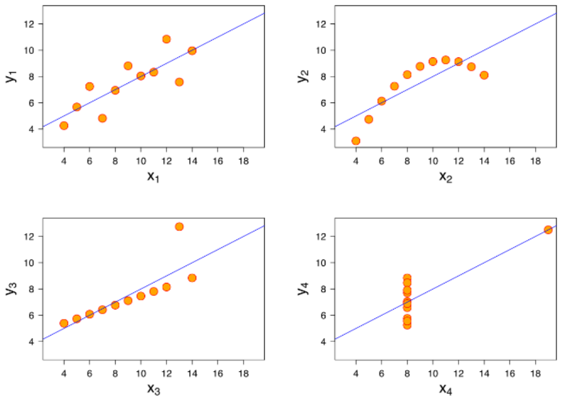
Violation of assumptions leads to unreliable inference. Also, linear regression is used to interpolate the data within the range. It cannot be used for extrapolation.

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>

There are 3 lies – lies, damned lies and statistics.
Anscombe's quartet will confirm the above line and it will be used to illustrate the importance of data visualization and limitations of statistics summary. It will emphasize the relevance of exploratory data analysis to spot the outliers, trends, and other hidden patterns which cannot be spotted in the summary statistics.

Anscombe's quartet is a set of 4 datasets having identical summary statistics in terms of standard deviation, variance, mean, correlation etc. By referring to this summary statistics, we can assume or infer that all the datasets are behaving identical. So, 1 solution for all. This is wrong inference. Anscombe's quartet emphasize the data visualization. Let's visualize those data set, you can see the different patterns.



We can see 4 different patterns from the 4 different datasets even though the summary statistics is same. This phenomenon will assert the importance of exploratory data analysis to infer the hidden patterns, trends and outliers. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading

This is called Anscombe's quartet. This exercise highlights how data distributions can yield the same statistical results, underscoring the value of exploratory data analysis of any data sets.

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>

Generally, correlation is a statistical measure to identify the relationship between the 2 variables. It is a vector based, which denotes the magnitude (strength) and direction of a relationship.

Pearson's R (Pearson correlation coefficient) is a type of correlation which measures the relationship between the continuous variables. It quantifies the relationship in the scale between -1 and 1.

- 1 – Positive correlation (positive slope)
- 0 – No correlation (no slope)
- -1 – Negative correlation (negative slope)

FORMULA:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

The Pearson R is used when the both the variables are quantitative (continuous), normally distributed, no outliers and the relationship is linear.

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 Probability, there is a concept called TRANSFORM RANDOM VARIABLES. In this, there are 2 concepts which is crucial – Scaling and Shifting. Both these methods are used to transform the random variable without affecting its properties and originality.

Scaling – It is used to scale all the variables to adjust on the same range without affecting its properties. It is used to apply the machine learning algorithm to converge faster and accurate results. There are two types of scaling

1. MinMax Scaler (normalization) – It scale the data within the range of 0 and 1. It is highly influenced by outlier.
2. Standard Scaler (Standardization) – It transforms the data with mean 0 and standard deviation of 1. It is robust as it uses mean and standard deviation.

Deciding which scaling to use is based on the type of data we are using in the model. If the data is normally distributed whose central tendency is similar, then we can go for standard scalar. Whereas, the min-max scaler is used for bounded data where there is no outliers. Using these 2 methods, will not affect the predictive power.

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>

Variance Inflation Factor is generally used to identify the multicollinearity between the variables. The VIF will show how much variance it brings. When VIF becomes infinite, it denotes that the variable is highly correlated (perfect correlation) or perfect multicollinearity, redundant features and makes the model complicated to perform matrix operation behind the linear regression.

Once we identified the VIF with high value or infinite, then we need to remove it to avoid the redundancy and complication.

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 (Quantile-Quantile plot) is a graphical representation of data to identify whether the data is normally distributed. It points the dataset that split the data into intervals of equal proportions.

1. Quartile (25%, 50% and 70%)
2. Median (50%)
3. Percentiles (1-100%)

It plots the quantiles of the first data set against the quantiles of the second data set. Need to create a scatter plot with the observed quantiles on the x-axis and theoretical quantiles on the y-axis. Theoretical quantiles means that quantiles calculated from theoretical gaussian distribution. It is widely used in linear regression to provide the normality of residuals to validate the assumptions and improve the reliability of regression results.