Assignment-based Subjective Questions

Q1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans) There were 6 categorical variables in the dataset.

We used Box plot (refer the fig above) to study their effect on the dependent variable ('cnt') .

The inference that We could derive were:

1. **season**: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.
2. **mnth**: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
3. **weathersit**: Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
4. **holiday**: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.
5. **weekday**: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.
6. **workingday**: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

---

Q2) Why is it important to use drop_first=True during dummy variable creation?

Ans) drop_first=True is important to use, as **it helps in reducing the extra column created during dummy variable creation**. Hence it reduces the correlations created among dummy variables.

Q3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans) The Pair-Plot tells us that there is a LINEAR RELATION between 'temp','atemp' and 'cnt'. It seems that both 'temp' & 'atemp' are equally correlated with target variable ‘cnt’ and thus they have the maximum correlation with the target variable.

Q4) How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans)
1. There should be a linear relation between independent variable and the target variable. To ensure this we will see the pair plot.

2. There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation. To ensure this we plot a residual plot.

3. The independent variables must not be correlated (no multicollinearity). To ensure this we check the VIF (variable inflation factor) for the independent variables. It should be less than 5.

4. The error terms must have constant variance (homoscedastic). To ensure this we can check the residual plot.

5. Error terms must be normally distributed. We can plot a histogeam of residuals to check this.

Q5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans) As per our final Model, the top 3 predictor variables that influences the bike booking are:

- **Temperature (temp)** - A coefficient value of ‘0.5749’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.5749 units.

- **Weather Situation 3 (weathersit_3)** - A coefficient value of '-0.3094' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3094 units.
- **Year (yr)** - A coefficient value of '0.2304' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2304 units.

General Subjective Questions

Q1) Explain the linear regression algorithm in detail?

Ans) It is one of the most well understood algorithms in statistics and machine learning,

Linear regression is an attractive model because the representation is so simple.
The representation is a linear equation that combines a specific set of input values (X) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

$$y = mX + b$$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slop of the regression line which represents the effect X has on Y

b is a constant, known as the Y-intercept. If X = 0,Y would be equal to b.

Simple Linear Regression

It is a model to describe the relationship between single dependent variable y and single independent variable x

$y = \beta_0 + \beta_1 x + c$

Multiple Linear Regression

It's a form of linear regression that is used when there are two or more predictors.

$y = \beta_0 + \beta_1 x + \beta_2 x + ... + \beta_p x_p$

where *Y* is the output variable, and *X* terms are the corresponding input variables and each predictor has a corresponding slope coefficient (*β)*

Steps to be followed in Linear Regression :-

- Reading and understanding the data
- Visualizing the data (Exploratory Data Analysis)
- Data Preparation
- Splitting the data into training and test sets
- Building a linear model
- Residual Analysis of the train data
- Making predictions using the final model and evaluation

Q2) Explain the Anscombe's quartet in detail.

Ans)

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Using only descriptive statistics and found that the mean, standard deviation, and correlation between x and y.

Anscombe's Quartet reminds that graphing data prior to analysis is good practice, outliers should be removed when analyzing data, and statistics about a data set do not fully depict the data set in its entirety.

Q3) What is Pearson's R?

Ans) Pearson correlation coefficient also known as Pearson's r, it is measure of linear correlation between two sets of data and the result always has a value between −1 and 1. A -1 means there is a strong negative correlation and +1 means that there is a strong positive correlation. A 0 means that there is no correlation (this is also called zero correlation).

Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- Pearson's Correlation Coefficient formula is as follows,

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

where,
r = Pearson Coefficient
n= number of the pairs of the stock
∑xy = sum of products of the paired stocks
∑x = sum of the x scores
∑y= sum of the y scores
∑x2 = sum of the squared x scores
∑y2 = sum of the squared y scores

Q4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans) Scaling is a method used to normalize the range of independent variables or features of data, it is also known as data normalization and is generally performed during the data preprocessing step.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1.

- sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

MinMax Scaling :x = x- min(x)/max(x)-min(x)

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

- sklearn.preprocessing.scale helps to implement standardization in python.

- Disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Standardization: x = x- mean(x)/sd(x)

- Standardization may be used when data represent Gaussian Distribution, while Normalization is great with Non-Gaussian Distribution
- Impact of Outliers is very high in Normalization

Q5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans) VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Q6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans) It is known as Quantile-Quantile plots, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution.

Quantiles are breakpoints that divide our numerically ordered data into equally proportioned buckets. Quartiles are quantiles that divide our data into 4 buckets (0–25%, 25–50%, 50–75%, 75–100%).

It is to determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.